Stock Price Prediction with Sentiment Analysis for Chinese Market

Yuchen Luan[1], **Haiyang Zhang**[1*], **Chenlei Zhang**[1], **Yida Mu**[2], **Wei Wang**[1]
[1]Xi'an Jiaotong Liverpool University
[2]The University of Sheffield
{yuchen.luan22, chenlei.zhang}@student.xjtlu.edu.cn
{haiyang.zhang, wei.wang03}@xjtlu.edu.cn
y.mu@sheffield.ac.uk

Abstract

Accurate prediction of stock prices is considered as a significant practical challenge and has been a longstanding topic of debate within the economic domain. In recent years, sentiment analysis on social media comments has been considered an important data source for stock prediction. However, most of these works focus on exploring stocks with high market values or from specific industries. The extent to which sentiments affect a broader range of stocks and their overall performance remains uncertain. In this paper, we study the influence of sentiment analysis on stock price prediction with respect to (1) different market value groups and (2) different Book-to-Market ratio groups in the Chinese stock market. To this end, we create a new dataset that consists of 24 stocks across different market value groups and Book-to-Market ratio categories, along with 12,000 associated comments that have been collected and manually annotated. We then utilized this dataset to train a variety of sentiment classifiers, which were subsequently integrated into sequential neural-based models for stock price prediction. Experimental findings indicate that while sentiment integration generally improve the predictive performance for price prediction, it may not consistently lead to better results for individual stocks. Moreover, these outcomes are notably influenced by varying market values and Book-to-Market ratios, with stocks of higher market values and B/M ratios often exhibiting more accurate predictions. Among all the models tested, the Bi-LSTM model incorporated with the sentiment analysis, achieves the best prediction performance.

Keywords: Stock Price Prediction, Sentiment Analysis, Chinese Stock Market

1. Introduction

Stocks are frequently traded investment products, and accurately forecasting stock prices is regarded as a crucial practical concern. This topic has been a subject of ongoing debate in the field of economics, with numerous scholars proposing various methods to forecast stock market trends. In recent years, the rise of social media has led many investors to express their views and sentiments on stocks in online forums, prompting scholars and practitioners to pay attention to discourse on these investment platforms. Such information has been shown to offer evidence indicating that investor sentiment might play a pivotal role in explaining stock price fluctuations (Dewally, 2003; Sunny et al., 2020).

Most existing works on stock prediction with sentiment analysis follow a two-stage process: the first stage involves using sentiment classification methods to compute sentiment values, which are subsequently integrated into conventional time series stock price prediction models. (Jing et al., 2021; Tashiro et al., 2019; Sirignano and Cont, 2021; Hiew et al., 2019; Sidogi et al., 2021). Common models employed for sentiment analysis include Convolutional Neural Networks (CNNs) and BERT models. Sentiment analysis often employs various models, including Convolutional Neural Networks (CNNs) and BERT-based models, to interpret and classify emotions within text data effectively. Specifically, when analyzing sentiment in Chinese text, a significant number of studies prefer the Bert-base-Chinese model (BBC) for its general applicability. However, a smaller yet noteworthy body of research opts for the Erlangshen-MegatronBert-1.3B-Sentiment model (EMB-1.3B-S), which has been shown to outperform others in classification tasks, as highlighted in the literature (Zhang et al., 2022). As for the stock prediction task, the majority of studies aim to predict the future direction of stock movements as a classification task. In contrast, a lesser-explored avenue is to predict the exact stock price based on historical data, treating it as a regression task. For this latter task, Long-Short Term Memory (LSTM) networks are frequently chosen due to their proficiency in processing and analyzing time series data (Hiew et al., 2019; Sidogi et al., 2021).

In the realm of stock prediction research, a prevalent trend involves selecting stocks based on criteria such as market capitalization (Zhang et al., 2017; Liu and Chen, 2019) or industry sector (Huang et al., 2018; Wu et al., 2018). However, such methods introduce a selection bias where the chosen stocks often share similar features, leading to a lack of diversity within the analyzed portfolio. Even when

* denotes corresponding author.

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 167–177

considering both market capitalization and industry factors together, it remains challenging to avoid the concentration of market capitalization within specific industries (Jing et al., 2021). For instance, stocks in the banking and food and beverage industries typically have high market capitalization, while those in the chemical and communication equipment industries tend to have lower market capitalization. This leads to an issue of similarity among the stocks to be predicted within the portfolio. An innovative approach to counteract this bias involves incorporating the Book-to-Market (B/M) ratio (Pontiff and Schall, 1998), a pivotal metric in value investment strategies indicating company valuation. Considering both the B/M ratio and market capitalization for stock selection can effectively mitigate this selection bias. In this paper, we examine the influence of sentiment analysis on stock price prediction with respect to (1) different market value groups and (2) different Book-to-Market ratio groups in the Chinese stock market. We train a set of sentiment classifiers, which are then incorporated with sequence-based deep learning models for price prediction. The contributions of this work are as follows:

- We construct a new dataset comprising 24 stocks from various market value and book-to-market ratio groups in the Chinese stock market, along with 12,000 corresponding comments that were collected and manually annotated.
- We employ various combinations of sentiment analysis models and sequence-based price prediction models to assess the impact of sentiment information on stock prediction.
- Experimental results suggest that while incorporating sentiment generally improves predictive performance, it may not consistently lead to superior results for individual stocks. Furthermore, the results are significantly influenced by different market values and Book-to-Market ratios. Among all the models tested, the Bi-LSTM model integrated with a sentiment factor demonstrates the highest prediction performance.

2. Datasets

2.1. Stock Selection

Considering the diverse market attributes of stocks in different market value portfolios in the Chinese market, we selected stocks from four market indexes from the Shanghai Stock Exchange (SSE)[1], namely `CSI 100`, `CSI 200`, `CSI 500`, and `CSI 1000`, representing portfolios of stocks with different market capitalizations and liquidity in the Chinese stock market. The `CSI 100` comprises the top 100 stocks with the largest market capitalization and best liquidity from the Shanghai and Shenzhen 300 indices, representing mega-cap stocks in the Chinese market; the `CSI 200` consists of 200 stocks excluding the constituents of the `CSI 100` index, representing large-cap stocks; the `CSI 500` and `CSI 1000` represent mid-cap and small-cap stocks, respectively. Subsequently, we constructed a 3×4 table by combining the three market capitalization portfolios with four B/M ratio portfolios. Six stocks meeting the selection criteria were randomly chosen from each cell of the table. For the 24 selected stocks, technical indicators including the opening price, closing price, highest price, lowest price, and trading volume have been collected from the China Stock Market & Accounting Research Database (CSMAR)[2]. Regarding technical indicators, we employ the Lagrange interpolation method to rectify missing and outlier values, subsequently arranging the data chronologically (de Resende et al., 2016).

Time Span To capture highly diverse price fluctuations and to alleviate concerns about data snooping, we selected data spanning from January 1, 2017, to December 31, 2022, covering 1,459 trading days. This interval has been deemed adequate by prior research for stock price prediction purposes, capturing essential fluctuations in market sentiment (Jiang, 2021). This selection ensures a comprehensive analysis period that incorporates significant market events and trends, providing a robust foundation for evaluating the impact of market sentiment on stock price movements.

2.2. Stock Comments Collection

For the experiments, we collected over 1.2 million stock comments related to the 24 selected stocks from the stock forum on the Financial Website (East Money)[3] for the corresponding 24 stocks. Given East Money's reputation as a leading financial information platform in China, the discussions on this forum are indicative of the broader sentiment among Chinese investors (Wang et al., 2018).

Data Filtering To ensure adherence to the fundamental requirements and standards of this experiment, we systematically excluded stocks previously categorized under *ST* or **ST* status[4], elimi-

[1] http://english.sse.com.cn/

[2] http://www.data.csmar.com

[3] http://www.guba.eastmoney.com

[4] In the Chinese stock market, *ST* represents "Special Treatment," indicating companies facing the risk of delisting due to financial distress or other issues, while

Average B/M \ Market Value	CSI 100	CSI 200	CSI 500	CSI 1000
High (33%)	601818.SH 601998.SH	000783.SH 600741.SH	000488.SH 600657.SH	000797.SH 601588.SH
Medium (33%)	600999.SH 002736.SH	600085.SH 601021.SH	000685.SH 300244.SH	002138.SH 002542.SH
Low (33%)	600585.SH 601336.SH	300144.SH 300033.SH	603355.SH 600259.SH	603989.SH 300377.SH

Table 1: Stocks selected based on the Fama-French three-factor model.

nated those with less than 50 trading weeks annually, and removed entries marked by missing data. This process retained stocks that consistently maintained their status as index component constituents throughout the designated trading period. The selected 24 stocks are listed in Table 1.

Text Pre-processing For the stock comments, we remove all garbled text, web links, and irrelevant short phrases.

Data Annotation Three annotators independently assigned sentiment labels: positive, neutral, or negative, to a set of 12,000 stock comments. To evaluate the inter-annotator reliability concerning the sentiment classification of these comments, we employed the Fleiss' Kappa (Fleiss, 1971) statistic as our chosen metric. This approach facilitates a quantifiable assessment of agreement levels among the three annotators, ensuring the consistency and accuracy of the sentiment labels assigned to the dataset. The Fleiss' Kappa is computed as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

where $p_o = \frac{1}{N}\sum_{i=1}^{N} p_i$ is the average observed agreement probability across all raters for all samples, and $p_i = \frac{1}{n(n-1)}\sum_{j=1}^{k} n_{ij}(n_{ij}-1)$represents the degree of agreement observed among raters for each sample. n_{ij} is the number of raters who classified sample i into category j, n is the total number of raters (in this study, there are 3 raters), and k is the number of categories (in this study, there are 3 categories: positive, negative, and neutral). $p_e = \sum_{j=1}^{k} p_j^2$ represents the expected average agreement probability that raters can achieve when assigning ratings. $p_j = \frac{1}{Nn}\sum_{i=1}^{N} n_{ij}$ represents the mean number of raters assigned to each category, where N is the total number of samples. The Fleiss Kappa value for our annotation is **0.883**, indicating excellent agreement and demonstrating good classification consistency.

*ST denotes a more severe level of "Special Treatment."

3. Methodology

We propose a hybrid predictive pipeline that combines 1) a sentiment analysis model to predict the sentiment score based on the daily comments for each stock, and 2) a sequence model to predict time series stock price that includes the sentiment factor. The architecture of the proposed method is depicted in Figure 1.

3.1. Sentiment Analysis on Stock Comments

We explore a number of text classification methods for predicting the sentiment of stock comments, including traditional machine learning models (e.g., Support Vector Machine (SVM)) and neural-based models, such as Convolutional Neural Networks (CNN) (Luan and Lin, 2019) and Transformer-based models (Vaswani et al., 2017; Devlin et al., 2019).

SVM SVM is used as a baseline for our sentiment classification. It utilizes unigram and bigram bag-of-words, weighted using TF-IDF, as inputs. These are implemented using the default settings of scikit-learn (Pedregosa et al., 2011).

CNN CNN arpproaches leverage multiple convolutional kernels of varied granularities to meticulously extract text features. . This process begins with the generation of feature matrices, followed by the execution of one-dimensional convolution and pooling operations to distill and condense the information. The culmination of this process involves the application of the Softmax function for sentiment classification, which computes a probability distribution across the possible sentiment categories for a given text. Following (Kim, 2014),our approach integrates pre-trained word embeddings through two distinct embedding layers: static and non-static. The filter size is set to 3, where each type of filter comprises 100 filters. Then, max-pooling operations are employed to extract critical information, ultimately yielding output results in the fully-connected layer.

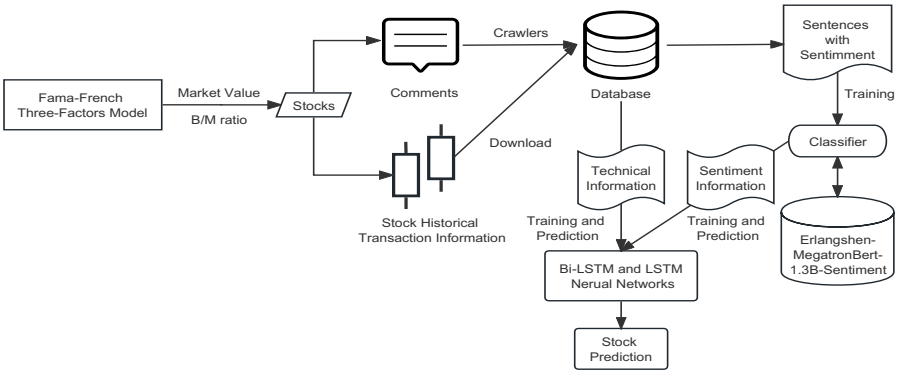


Figure 1: The design of the stock price prediction model in this study based on sentiment analysis.

CBERT & EMB-1.3B-S BERT (Devlin et al., 2019; Zhang et al., 2022), the pre-trained deep bidirectional Transformer, has shown strong performance on many NLP tasks (Devlin et al., 2019). Conventionally, it is pre-trained using two self-supervised tasks (masked language modeling and next sentence prediction) on a large corpus and fine-tuned for downstream tasks. In this paper, we fine-tune two pre-trained BERT models for Chinese for the sentiment classification task: Chinese Bidirectional Encoder Transformers[5] (CBERT)(Cui et al., 2021) and Erlangshen-MegatronBert-1.3B-Sentiment[6] (EMB-1.3B-S).

CBERT is pre-trained on an extra Chinese corpus (e.g., news articles and social media posts), based on the pre-existing checkpoint of the BertBase-Chinese model (Devlin et al., 2019), maintaining an identical structure (e.g., 12 layers and 110M parameters) to the vanilla BERT-base model. It achieves comparable predictive performance on multiple Chinese NLP downstream tasks compared to traditional machine learning approaches. EMB-1.3B-S, one of the largest open-source Chinese BERT models to date with 1.3 billion parameters, surpasses human performance on downstream tasks such as the TNEWS[7] Subtask.

We employ CBERT and EMB-1.3B-S in our task by incorporating an additional linear layer on top of the 12-layer transformer blocks with a Sigmoid activation, following the standard model fine-tuning pipeline introduced by (Devlin et al., 2019). For both transformer-based models, we set the maximum input length to 512 tokens. Additionally, to maintain consistency with the time input of the stock price prediction model, we computed the daily sentiment value (SV_t) for each trading day using the following equation:

$$SV_t = \frac{num_t^+ \cdot \mathrm{TScores}_t^+ - num_t^- \cdot \mathrm{TScores}_t^-}{num_t} \quad (2)$$

where $\mathrm{TScores}_t^+$ and $\mathrm{TScores}_t^-$ represent the sum of sentiment probability scores for all positive and negative labels corresponding to a stock on the t-th trading day, respectively; num_t denotes the total number of comments on the t-th trading day. The sentiment value ranges from -1 to 1, indicating the overall investor sentiment towards a particular stock on that day: a positive value suggests a predominance of positive sentiments, and a negative value indicates the opposite.

3.2. Stock Prediction with Sentiment Analysis

In our study, we deploy both Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (Bi-LSTM) models, synergistically combined with sentiment analysis, to forecast stock prices. Specifically, we adopt a sliding window technique for predicting stock prices for the subsequent day. This method involves progressively moving the input window over the data series to generate predictions for each new time step. This approach allows for dynamic analysis of time-series data, where the LSTM and Bi-LSTM frameworks leverage both historical stock prices and sentiment indicators within each window to make informed predictions about future stock price movements.

LSTM Long Short-Term Memory (LSTM) networks, a subclass of recurrent neural networks (RNNs), enhance the RNN framework by effectively managing sequential data while overcoming the notorious gradient vanishing and exploding issues commonly associated with traditional RNNs (Hochreiter and Schmidhuber, 1997). LSTMs introduce a unique mechanism for long-term memory retention, enabling the model to make judicious

[5] https://huggingface.co/hfl/chinese-bert-wwm

[6] https://huggingface.co/IDEA-CCNL/Erlangshen-TCBert-1.3B-Sentence-Embedding-Chinese

[7] Toutiao News Classification Dataset

use of relevant historical information without being overly dependent on distant past data. This feature ensures a more balanced consideration of both recent and older inputs, significantly improving the network's ability to learn from sequences over extended periods.

The calcluation of the LSTM are is shown as follows:

$$\begin{cases} i_t = \sigma(w_i \cdot [H_{t-1}, X_t] + b_i) \\ f_t = \sigma(w_f \cdot [H_{t-1}, X_t] + b_f) \\ \tilde{C}_t = \tanh(w_c \cdot [H_{t-1}, X_t] + b_c) \\ o_t = \sigma(w_o \cdot [H_{t-1}, X_t] + b_o) \\ C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \\ h_t = o_t \cdot \tanh(C_t) \end{cases} \tag{3}$$

where t represents the time point, X_t signifies the input value at the cell, and H_t represents the output state of the cell at the same time point. The symbols f_t, i_t, and o_t correspond to the formulas for the forget gate, input gate, and output gate, respectively. C_t denotes the cell state update. Matrices w_i, w_f, w_c, and w_o are the weight matrices for the input gate, forget gate, update gate, and output gate, respectively. Biases b_i, b_f, b_c, and b_o represent the respective biases. The activation function σ is applied to each gate unit, generating values between 0 and 1. This activation function is also applied to the cell state and output, constraining their values to a range between -1 and 1.

To enhance the LSTM model's capability for stock price prediction, we integrate sentiment values as supplementary features. More precisely, we concatenate the sentiment value SV_t as an additional feature of the input data, forming an augmented input vector, as shown in Equation 4.

$$i_t = \sigma(w_i \cdot [H_{t-1}, X_t, SV_t] + b_i) \tag{4}$$

By incorporating these sentiment values, they directly influence the operations of the input gate, forget gate, and the calculation of the input candidate value. This strategic integration empowers the model to adeptly leverage sentiment information, refining its ability to predict stock prices by learning from the nuanced interplay between market sentiment and stock price movements during the training phase.

Bi-LSTM The Bi-LSTM model, initially proposed by (Graves and Schmidhuber, 2005), consists of two LSTM layers that enable bidirectional processing of stock price information around time t. By leveraging historical data from both forward and backward directions, it jointly predicts the stock's closing price at time t. Bi-LSTM structure consists of two distinct LSTM layers aligned in parallel, each processing the temporal data sequence in opposite directions: one forward and the other backward. This setup allows for the comprehensive assimilation of contextual information, both preceding and following the target time t, thereby enriching the model's understanding and predictive accuracy of stock price movements by leveraging insights from both past and future contexts. The calculation of Bi-LSTM is represented as:

$$\begin{cases} \overrightarrow{h_t} = LSTM(\overrightarrow{h_{t-1}}, x_t) \\ \overleftarrow{h_t} = LSTM(\overleftarrow{h_{t+1}}, x_t) \\ h_t = (\overrightarrow{h_t}, \overleftarrow{h_t}) \end{cases} \tag{5}$$

Similarly, we integrate sentiment factors into the computation, where at each time step t, the sentiment factor is included in Bi-LSTM as part of the input x_t:

$$h_t = LSTM(h_{t-1}, x_t, SV_t) \tag{6}$$

This approach allows sentiment factors to influence the input gate, forget gate, and input candidate value computations, enabling the model to learn how to effectively use sentiment information for stock price prediction during the training process.

4. Experiment and Results

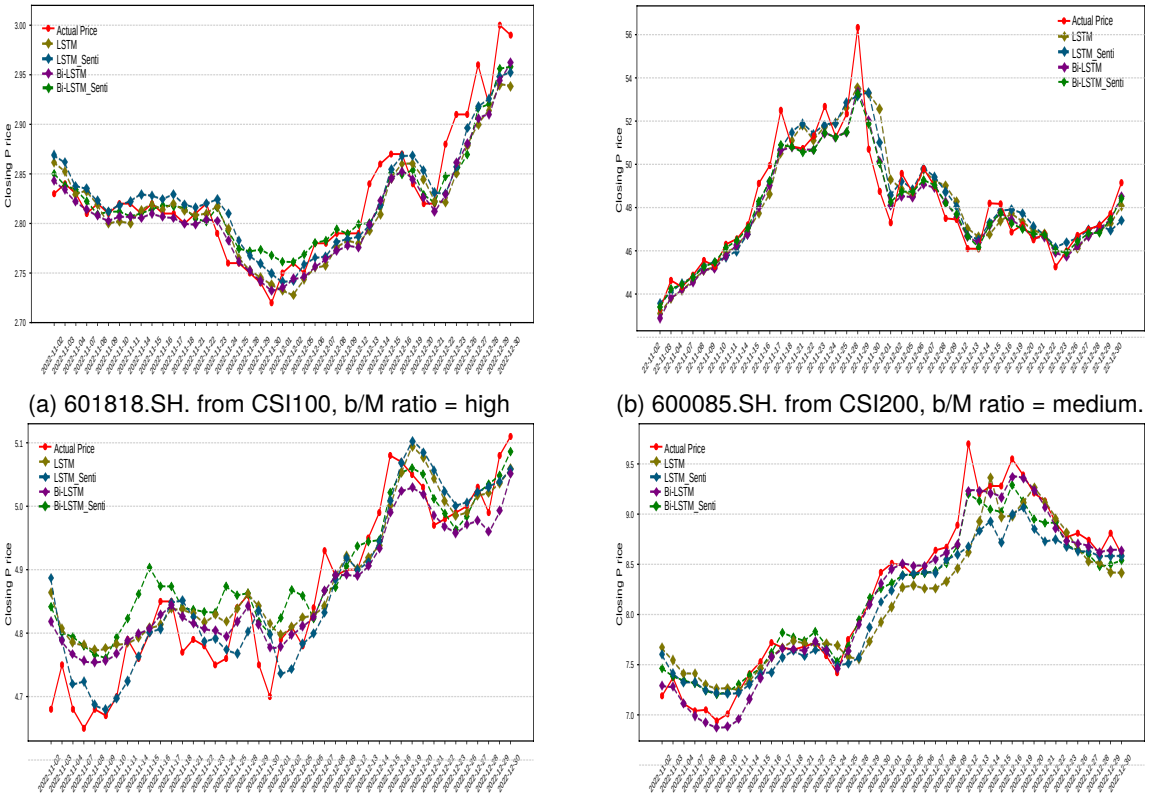
4.1. Experiments on Sentiment Analysis

To evaluate the predictive performance of various classifiers on sentiment classification, we use comments collected from January 1, 2017, to October 30, 2022, as the training set, and comments from November and December 2022 as the test set. We report precision, recall, and the F1 measure to assess their performance.

The average evaluation results are presented in Table 2. Considering the presence of data imbalance within the dataset, we employ a micro-average method for calculating the F-measure. As indicated in Table 2, the EMB-1.3B-S model achieves the best overall performance. Given that the number of comments collected from forums exceeds 1.2 million, this level of improvement can significantly enhance the accuracy of sentiment judgment. Therefore, employing this classifier for analyzing the hidden sentiments in text data collected from forums is feasible. In the stock price prediction phrase, we utilize the results from the EMB-1.3B-S model as one of the input features.

Metric	Precision	Recall	F-1
SVM	0.823	0.764	0.792
CNN	0.875	0.823	0.848
CBERT	0.947	0.946	0.946
EMB-1.3B-S	**0.970**	**0.969**	**0.969**

Table 2: Performance Comparison of Various Classifiers in Sentiment Analysis.



(a) 601818.SH. from CSI100, b/M ratio = high

(b) 600085.SH. from CSI200, b/M ratio = medium.

(c) 000488.SH. from CSI500, b/M ratio = high

(d) 300377.SH. from CSI1000, b/M ratio = low.

Figure 2: The actual closing price and the predicted prices across the four model combinations on one stock from each stock market index.

4.2. Experiments on Stock Price Prediction

In our experimental setup for forecasting stock prices, we merge sentiment scores obtained from sentiment analysis with technical indicators related to the stock market to predict the closing prices for the following day. This integration approach combines qualitative insights from investor sentiment with quantitative stock technical factors, providing a comprehensive view that enhances the accuracy of our predictive model for next-day closing prices. Same train/test data split are used as that of sentiment analysis. Two metrics are employed: Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) to evaluate the performance. Smaller values of these two metrics indicate that the model's predictions are closer to the actual values.

For both LSTM and Bi-LSTM, we set the input length to 3, and use 64D-3 layer neural networks. The batch size is 32. Figure 2 demonstrates the performance comparison using different methods: 1) LSTM, 2) LSTM with sentiment factor, 3) Bi-LSTM and 4) Bi-LSTM with sentiment factor, against the actual stock prices for one stock (600818.SH, 600085.SH, 000488.SH, 300377.SH) from each market index. It demonstrates that all prediction models accurately forecast the stock price trends.

Influence of Sentiment Factor Table 3 presents the aggregated performance of stocks from different market value groups using various models, with the best performance highlighted in bold. Performance for individual stock within each market value group are provided in Appendices. It is observed that incorporating sentiment information does not uniformly enhance prediction accuracy for every stock. This observation suggests that the effectiveness of sentiment data integration varies across different stocks, indicating a nuanced relationship between sentiment analysis and stock performance forecasting.

Table 4 aggregates the performance metrics for all stocks analyzed through various models, highlighting the comparative results. Notably, the Bi-LSTM model, augmented with sentiment data, demonstrates the best results, achieving a RMSE of 41.1603 and a MAPEof 145.5350. In contrast, the LSTM model that does not incorporate sentiment factors registers the least favorable outcomes, with an RMSE of 41.3073 and a MAPE of 148.6382. These findings indicate that integrating

sentiment information for stock prediction can generated overall better performance.

Performance on different Market Value and B/M ratio groups Furthermore, by segmenting the results according to various market capitalizations and Book-to-Market (B/M) ratios, we noted marked variations in model performance across different segments, as detailed in Table 5. Particularly, the `CSI 100` group exhibited the best performance, with an RMSE of 24.4222 and a MAPE of 102.4448. Conversely, the `CSI 200` group recorded the highest RMSE at 73.8258, while the `CSI 1000` group had the highest MAPE at 186.7908, indicating that the model performs excellently in predicting stocks with higher market values.

We also conducted an analysis to evaluate the impact of sentiment factor on different Book-to-Market (B/M) ratios, with the results detailed in Table 6. The findings indicate that stocks categorized within the High B/M ratio group exhibited the most accurate predictions, with an RMSE around 3.8, showcasing their robustness in predictive accuracy. In contrast, stocks within the Low B/M ratio group displayed the least favorable performance. It also reveals a trend where the overall RMSE progressively increases as the B/M ratio shifts from High to Low.

4.3. Ablation Study

Influence of B/M ratio To assess the influence of the Book-to-Market (B/M) ratio and the effect of integrating sentiment analysis on the models' overall efficacy, we embarked on a detailed ablation study. Specifically, we investigated the relationship between daily sentiment values and actual closing prices for stocks grouped by their B/M ratios. For this purpose, we employed the Pearson correlation coefficient (Asuero et al., 2006) as our primary metric. This coefficient is determined by:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{7}$$

Table 7 demonstrates the Pearson correlation coefficients between sentiment scores and closing prices within different B/M ratio groups. It is observed that the highest correlation is observed within the high B/M ratio group, suggesting a pronounced link between sentiment scores and closing prices in this group. This finding aligns with the superior performance of our predictive model within the high B/M ratio group.

Influence of market index We noted that the proposed model combination performed best in the CSI 100 portfolio, consistent with the characteristics of large-cap companies, which typically possess advantages such as high stability, high liquidity, and comprehensive information disclosure. High stability and liquidity often manifest as relatively stable technical indicators, favoring predictions from single models. Moreover, comprehensive information disclosure implies richer information about high market value stocks, making them focal points for investors' attention, naturally accompanied by more stock comments. To verify this, We also investigated the relationship between the number of stock reviews and predictive results within different market value groups. It was observed that in the `CSI 100` high market value group, there were the most stock reviews (280,583 records), while in the `CSI 1000` low market value group, there were the fewest stock reviews (203,526 records). This finding aligns with the focus of public attention, as stocks with higher market values are typically associated with larger companies and enjoy greater exposure, thus attracting more stock review information.

Drawing from the insights garnered in this study, investors and analysts looking to leverage time series models for forecasting stock prices in the Chinese market might benefit from focusing on stocks characterized by high Book-to-Market (B/M) ratios and exceptionally large market values, specifically those within the CSI 100 category. These segments have shown to yield more accurate predictive outcomes. Additionally, for models that incorporate sentiment analysis into the stock price forecasting process, the Bi-Long Short-Term Memory (Bi-LSTM) model emerges as a more effective option compared to the Long Short-Term Memory (LSTM) model. This recommendation is based on the Bi-LSTM model's superior performance, especially when analyzing stocks with high B/M ratios, where the integration of sentiment factors enhances prediction accuracy.

5. Conclusion and Future Work

This study introduces a novel hybrid model for stock price prediction, alongside the creation of a comprehensive Chinese stock sentiment classification dataset. Experimental results show that the performance of machine learning models on stock prediction varies on different market index groups , with best performance on high market values (CSI 100). It also suggest that the integration of sentiment analysis into stock price prediction models generally leads to improved accuracy, although the extent of this improvement varies. The impact of incorporating sentiment analysis is not uniform across all stocks, with noticeable differences based on market value and Book-to-Market (B/M) ratio segments and different market index groups. Intriguingly, for some stocks, the addition of sentiment data has been observed to diminish predictive performance, with such effects being especially marked within the low B/M ratio category.

Market Ind.	eval.	LSTM		Bi-LSTM	
		without	with senti	without	with senti
CSI 100	Total RMSE	6.1054	6.1112	6.1156	**6.0800**
	Total MAPE	25.4933	25.5460	25.8655	**25.4900**
CSI 200	Total RMSE	18.8044	18.2662	**18.2409**	18.5133
	Total MAPE	35.8019	**35.1904**	35.7643	35.9698
CSI 500	Total RMSE	9.5629	**9.3874**	9.5060	9.4294
	Total MAPE	39.5894	**38.9313**	39.3167	39.1492
CSI 1000	Total RMSE	7.3246	7.3893	7.3096	**7.0375**
	Total MAPE	48.7534	47.5120	46.1122	**44.4132**

Table 3: Aggregated performance for stocks from each market index group using different models.

Combination	LSTM	LSTM with senti	Bi-LSTM	Bi-LSTM with senti
Total RMSE	41.3073	41.1638	41.2541	**41.1603**
Total MAPE	148.6382	147.1796	147.0567	**145.5350**

Table 4: The total RMSE and MAPE aggregated by the combined predictive model.

Market value	CSI 100	CSI 200	CSI 500	CSI 1000
Total RMSE	**24.4222**	73.8258	37.8857	28.0510
Total MAPE	**102.4448**	142.7264	157.9866	186.7908

Table 5: Total RMSE and MAPE aggregated by market value.

B/M	LSTM			LSTM with senti		
	High	Medium	Low	High	Medium	Low
Total RMSE	3.8537	13.3996	24.0540	3.8124	12.8506	24.4802
Total MAPE	48.2670	43.9803	56.3907	47.2058	43.2707	56.7031

B/M	Bi-LSTM			Bi-LSTM with senti		
	High	Medium	Low	High	Medium	Low
Total RMSE	3.8493	13.4192	23.9037	3.7799	13.1928	24.2056
Total MAPE	46.5322	43.4357	57.0888	45.3580	43.9159	56.4611

Table 6: RMSE and MAPE for different models across different Book-to-Market ratio groups.

B/M ratio	High	Medium	Low
Pearson	**0.7104**	0.4705	0.1259

Table 7: The correlation coefficients between sentiment values and stock closing prices across different B/M ratio groups.

This research, while providing valuable insights, is subject to certain limitations. The predictive outcomes detailed in this study are derived solely from the context of the Chinese stock market and have not been tested across diverse market environments. The specific attributes of China's market, such as the absence of same-day buying and selling (T+0 trading), could potentially skew the applicability of our findings to other financial contexts. In our forthcoming efforts, we plan to broaden the scope of our investigation by integrating a wider array of sentiment analysis methodologies and including additional external market variables. This expansion aims to enhance the robustness and generalizability of our results, ensuring that our conclusions hold weight across varying global market dynamics.

6. Acknowledgements

We would like to acknowledge the support provided by the XJTLU AI University Research Centre, Jiangsu Province Engineering Research Centre of Data Science and Cognitive Computation at XJTLU and SIP AI innovation platform (No. YZCXPT2022103) and is also supported by the Research Development Funding (RDF) (No. RDF-21-02-044) at Xi'an Jiaotong-Liverpool University. Additionally, this research has received partial funding from the Jiangsu Science and Technology Programme (contract number BK20221260).

7. Bibliographical References

Agustin Garcia Asuero, Ana Sayago, and AG González. 2006. The correlation coefficient: An overview. *Critical reviews in analytical chemistry*, 36(1):41–59.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Damares CO de Resende, Ádamo Lima de Santana, and Fábio Manoel França Lobato. 2016. Time series imputation using genetic programming and lagrange interpolation. In *2016 5th Brazilian conference on intelligent systems (BRACIS)*, pages 169–174. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Michael Dewally. 2003. Internet investment advice: Investing with a rock of salt. *Financial Analysts Journal*, 59(4):65–77.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.

Joshua Zoen Git Hiew, Xin Huang, Hao Mou, Duan Li, Qi Wu, and Yabo Xu. 2019. Bert-based financial sentiment index and lstm-based stock return predictability. *arXiv preprint arXiv:1906.09024*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jieyun Huang, Yunjia Zhang, Jialai Zhang, and Xi Zhang. 2018. A tensor-based sub-mode coordinate algorithm for stock prediction. In *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*, pages 716–721. IEEE.

Weiwei Jiang. 2021. Applications of deep learning in stock market prediction: recent progress. *Expert Systems with Applications*, 184:115537.

Nan Jing, Zhao Wu, and Hefei Wang. 2021. A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. *Expert Systems with Applications*, 178:115019.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Jiexi Liu and Songcan Chen. 2019. Non-stationary multivariate time series prediction with selective recurrent neural networks. In *Pacific rim international conference on artificial intelligence*, pages 636–649. Springer.

Yuandong Luan and Shaofu Lin. 2019. Research on text classification based on cnn and lstm. In *2019 IEEE international conference on artificial intelligence and computer applications (ICAICA)*, pages 352–355. IEEE.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Jeffrey Pontiff and Lawrence D Schall. 1998. Book-to-market ratios as predictors of market returns. *Journal of financial economics*, 49(2):141–160.

Thendo Sidogi, Rendani Mbuvha, and Tshilidzi Marwala. 2021. Stock price prediction using sentiment analysis. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 46–51. IEEE.

Justin Sirignano and Rama Cont. 2021. Universal features of price formation in financial markets: perspectives from deep learning. In *Machine Learning and AI in Finance*, pages 5–15. Routledge.

Md Arif Istiake Sunny, Mirza Mohd Shahriar Maswood, and Abdullah G Alharbi. 2020. Deep learning-based stock price prediction using lstm and bi-directional lstm model. In *2020 2nd novel intelligent and leading emerging sciences conference (NILES)*, pages 87–92. IEEE.

Daigo Tashiro, Hiroyasu Matsushima, Kiyoshi Izumi, and Hiroki Sakaji. 2019. Encoding of high-frequency order information and prediction of

short-term stock price by deep learning. *Quantitative Finance*, 19(9):1499–1506.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Qili Wang, Wei Xu, and Han Zheng. 2018. Combining the wisdom of crowds and technical analysis for financial market prediction using deep random subspace ensembles. *Neurocomputing*, 299:51–61.

Huizhe Wu, Wei Zhang, Weiwei Shen, and Jun Wang. 2018. Hybrid deep sequential modeling for social text-driven stock prediction. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 1627–1630.

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, and Chongpei Chen. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.

Liheng Zhang, Charu Aggarwal, and Guo-Jun Qi. 2017. Stock price prediction via discovering multi-frequency trading patterns. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2141–2149.

A. Appendix

Stock ID	LSTM		LSTM with senti		Bi-LSTM		Bi-LSTM with senti	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
601818.SH	0.0723	2.0003	0.0644	1.7618	0.0717	1.9621	0.0597	1.6164
601998.SH	0.1733	3.1535	0.2213	3.5069	0.1948	3.4679	0.2287	3.6645
600999.SH	0.4904	3.0828	0.4966	3.1254	0.5142	3.2260	0.5008	3.1497
002736.SH	0.3146	2.8587	0.3118	2.8250	0.3113	2.8343	0.3041	2.7722
600585.SH	3.0708	8.6040	3.0450	8.5696	2.9508	8.3421	2.9175	8.2399
601336.SH	1.9840	5.7940	1.9721	5.7572	2.0728	6.0331	2.0792	6.0473

Table 8: RMSE and MAPE of the predicted results for stocks selected in CSI 100.

Stock ID	LSTM		LSTM with senti		Bi-LSTM		Bi-LSTM with senti	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
000783.SH	0.2051	3.0481	0.2018	3.0007	0.2098	3.1069	0.1990	2.9555
600741.SH	1.7092	7.5582	1.6745	7.3673	1.7176	7.5955	1.7031	7.5349
600085.SH	3.5355	5.8289	3.4708	5.7213	3.4208	5.4714	3.3259	5.6429
601021.SH	3.2210	5.0601	2.8503	4.5878	3.4136	5.3472	3.3309	5.2134
300144.SH	0.9565	6.2332	0.9268	6.0339	0.9763	6.3349	0.9507	6.1798
300033.SH	8.6771	8.0734	9.1420	8.4794	8.5029	7.9084	9.0037	8.4033

Table 9: RMSE and MAPE of the predicted results for stocks selected in CSI 200.

Stock ID	LSTM		LSTM with senti		Bi-LSTM		Bi-LSTM with senti	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
000488.SH	0.3031	4.9292	0.2988	4.8500	0.2899	4.7048	0.2989	4.8091
600657.SH	0.7743	12.3015	0.7668	12.1422	0.7774	12.3377	0.7628	12.0320
000685.SH	0.4096	4.8377	0.4098	4.8191	0.4163	4.8653	0.4263	5.0726
300244.SH	1.7188	4.6935	1.6800	4.5999	1.7034	4.6744	1.7232	4.6904
603355.SH	1.8167	4.4031	1.7429	4.2319	1.8246	4.4261	1.7819	4.3314
600259.SH	4.5404	8.4244	4.4691	8.2882	4.4944	8.3084	4.4364	8.2137

Table 10: RMSE and MAPE of the predicted results for stocks selected in CSI 500.

Stock ID	LSTM		LSTM with senti		Bi-LSTM		Bi-LSTM with senti	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
000797.SH	0.4968	10.7736	0.4784	10.4666	0.4758	9.0313	0.4157	8.6981
601588.SH	0.1196	4.5026	0.1064	4.1103	0.1123	4.3260	0.1040	3.8075
002138.SH	3.4924	12.1558	3.4245	11.9604	3.4415	12.0183	3.3665	11.7765
002542.SH	0.2173	5.4628	0.2274	5.6318	0.1981	4.9988	0.2051	5.0834
603989.SH	2.1708	6.5909	2.3174	7.0524	2.1710	6.6277	2.2018	6.6794
300377.SH	0.8377	8.2677	0.8352	8.2905	0.9109	9.1081	0.8344	8.3663

Table 11: RMSE and MAPE of the predicted results for stocks selected in CSI 1000.

B/M	LSTM			Bi-LSTM		
	High	Medium	Low	High	Medium	Low
Total RMSE	7.7366	26.2708	48.5342	7.6661	26.6120	48.1093
Total MAPE	95.4728	87.2510	113.0938	91.8902	87.3516	113.5499

Table 12: Total RMSE and MAPE for different models across different Book-to-Market ratio groups.