Expert Systems With Applications 210 (2022) 118312

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Predicting long-term stock movements with fused textual features of Chinese research reports

Ming Zhang [a,b], Jiahao Yang [a,b], Meilin Wan [c], Xuejun Zhang [a], Jun Zhou [a,*]

[a] Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, People's Republic of China
[b] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China
[c] Faculty of Physics and Electronic Technology, Hubei University, Wuhan 430064, People's Republic of China

ARTICLE INFO

Keywords:
Long-term stock prediction
Research report
Financial text mining
Textual feature
Pre-trained language model

ABSTRACT

By shaping investors' perceptions and assessments of the stock, research reports have significant impacts on the stock market. Due to the limitations of text mining technology, it is difficult for researchers to effectively utilize long research reports, and most studies mainly focus on investor sentiment. However, due to the lack of appropriate open-domain toolkits, the annotations of sentiment often require expensive manual labeling. In addition, most existing studies have shown the success of using textual data as a supplement to historical price data in short-term forecasting, but not in long-term forecasting. To cover this gap and solve the problem of difficult annotations, we introduce a novel knowledge-driven approach for long-term stock movement prediction based on Chinese research reports. In detail, a new long-term Stock Movement Prediction dataset composed of Research Reports is proposed, namely SMPRR. It is mainly composed of long, formal, and professional research reports and historical prices. Furthermore, we propose the Multi-module Feature Fusion method based on the pre-trained language model FinBERT (MFF-FinBERT), which can effectively fuse textual features from research reports. The experiment results show that the proposed model has achieved better performance than existing methods in the forecasting of one-year stock movements, and the accuracy reaches 79.2%. The results also indicate that the basic information of stocks plays an important role in long-term forecasting, which is in line with the theory of value investing.

1. Introduction

Stock movement prediction is widely considered difficult due to the volatile and non-stationary properties of the stock market (Sawhney, Agarwal, Wadhwa, & Shah, 2020; Zhang, Liu, & Zheng, 2021). However, the stock market always attracts investors seeking profitable returns because of its easy accessibility and high profitability (Yun, Yoon, & Won, 2021). Furthermore, the received wisdom of economics and finance holds that investors can predict stock movements by using publicly available data (Dumas, Kurshev, & Uppal, 2009). Over the decades, different techniques have been applied to model short- and long-term stock movement prediction, considering daily prices and other indicators from the stock market (Henrique, Sobreiro, & Kimura, 2019). Undoubtedly, the mechanism of stock movements is attracting more and more data scientists.

With the popularity of online trading platforms, more and more people are involved in the stock market (Yun et al., 2021). In recent years, long-term investing has been favored by junior investors due to lower transaction fees and easier management. Since long-term forecasting plays an important role in long-term investing, the research on high-precision long-term stock movement has greater research significance and application value. However, most of the existing studies tend to make daily or weekly forecasts, and it is difficult to carry out long-term stock movement predictions based on the modeling of historical prices.

During the last years, with the rapid development of natural language processing (NLP), research on predicting stock movements based on unstructured textual data has attracted considerable attention (Akita, Yoshihara, Matsubara, & Uehara, 2016). Most previous studies are based on tweets (Meesad & Li, 2014), news (Khedr & Yaseen, 2017), web posts (Zhao, Sun, & Cheng, 2020), and other information (Das & Chen, 2007), all of which are short, casual, and non-professional. However, the above textual data can only reflect part of the stock market. For example, online reviews can only represent the investor sentiment; financial news can only reflect the operating conditions of related companies. In general, it is difficult to dig out more powerful

* Corresponding author.
E-mail addresses: zhangming@hccl.ioa.ac.cn (M. Zhang), yangjiahao@hccl.ioa.ac.cn (J. Yang), wanml@hubu.edu.cn (M. Wan), zhangxuejun@hccl.ioa.ac.cn (X. Zhang), zhoujun@hccl.ioa.ac.cn (J. Zhou).

https://doi.org/10.1016/j.eswa.2022.118312
Received 17 April 2022; Received in revised form 20 July 2022; Accepted 26 July 2022
Available online 4 August 2022

market information from the above textual data for long-term stock movement prediction.

To solve the above problems, it is necessary to use long, formal, and professional textual data for long-term stock movement prediction. As well as known, research reports provide a valuable resource for harvesting professional analytic views on a company, providing deep insights and valuable conclusions. By shaping investors' perceptions and assessments of stock, research reports have significant impacts on the stock market (Daudert, 2021). The effectiveness of research reports in predicting stock movements can be measured from the following aspects. Firstly, stock price movements are considered relevant to research reports according to the inverse Efficient Market Hypothesis (EMH) (Roberts, 1959). Secondly, research reports written by professional analysts can provide a comprehensive overview of the company's business and financial condition (Yang, Zhang, & Fan, 2020). Thirdly, research reports have a better performance for one-year stock movement prediction based on the mathematical statistics in Section 4.2.2.

Due to the limitations of text mining technology, there are few studies using long textual data for long-term stock movement prediction. To the best of our knowledge, there are few datasets for long-term stock movement prediction that contain long, formal, and professional textual data. To bridge this gap, we make the following efforts to forecast one-year stock movements based on research reports.

First, we construct a large-scale long-term stock movement prediction dataset, named SMPRR. Specifically, it contains both historical prices and research reports. In detail, some preprocessing steps are taken to build SMPRR, such as using optical character recognition (OCR) technology to obtain the textual contents and relabeling them with historical prices. Each research report is divided into four modules based on themes, namely Facts-Opinions, Investment-Advice, Risk-Warnings, and Basic-Information modules. Moreover, SMPRR is thoroughly analyzed from three perspectives: label technology, forecast periods, and split methods. Then, we compare SMPRR with several widely-used open-source financial datasets.

Second, based on SMPRR, a family of strong and representative baselines is established. In detail, we first investigate the performance of several untrained models, including RAND, Research Report, and Market Trend. Furthermore, several classic non-pre-trained language models are implemented, including ARIMA (Ribeiro, Santos, Mariani, & Coelho, 2021), GRU (Wu, Wang, & Wu, 2022), LSTM (Zolfaghari & Gholami, 2021), Logistic Regression (Cakra & Trisedya, 2015), and Naïve Bayes (Feuerriegel & Gordon, 2018). Then, since pre-trained language models have been proven to be an effective way in stock movement prediction, several pre-trained language models are implemented, including BERT (Devlin, Chang, Lee, & Toutanova, 2019), RoBERTa (Liu, Ott, et al., 2019), and FinBERT (Liu, Huang, Huang, Li, & Zhao, 2020).

Third, upon these strong baselines, we further propose a multi-module feature fusion method based on FinBERT to better characterize the valuable contents of research reports, named MFF-FinBERT. The proposed method is mainly composed of three sub-components: multi-module feature extraction, multi-module feature fusion, and stock movement prediction. First, the predictive performance and the computational cost are compared. Moreover, we mainly discuss the importance of each module in research reports on long-term stock movement prediction. Then, an ablation study is performed to explore the effectiveness of multi-module feature extraction and multi-module feature fusion. Finally, a simulation of the profits obtained by using the proposed method is performed. The SMPRR dataset and the benchmark approaches can be obtained at https://github.com/zhangming-19/SMPRR.

The primary contributions of this work can be summarized as follows:

- We construct SMPRR, the long-term stock movement prediction dataset, which consists of 161,128 high-quality research reports composed of four sub-modules and the last ten years of historical prices. SMPRR will be released publicly for future research.
- We establish a series of neural baselines and provide benchmarks for long-term stock movement prediction using textual embeddings. Specifically, this work has successfully explored the prediction of long-term stock movements based on the textual features of research reports. Through various experiments, we observe consistent performance boosts originating from pre-trained language models, which verifies the significant merits of pre-training.
- We further propose a multi-module feature fusion method based on FinBERT, named MFF-FinBERT. By jointly processing the different modules, the proposed model can yield state-of-the-art results on SMPRR. This work has improved the accuracy of one-year stock movement prediction to 79.2%, surpassing the previous methods of 2.7%.

The rest of this paper is structured as follows. Related work and problem formulation are presented in Sections 2 and 3, respectively. Section 4 describes the construction and analysis of the proposed dataset. The details of the proposed model are described in Section 5. Then, Section 6 shows the experimental parts, and the experimental results are discussed in Section 7. Finally, Section 8 summarizes this work and plans for future work.

2. Related literature

2.1. Long-term stock movement prediction

Stock movement prediction is a challenging task due to the highly stochastic market and seriously chaotic data (Shah, Isah, & Zulkernine, 2019; Souza & Aste, 2019). The problem of long-term stock movement prediction is more complex than that of short-term movement prediction. Short-term forecasts mainly fall into technical analysis (TA), while long-term forecasts are based on fundamental analysis (FA). Traditional long-term works mainly focus on the modeling of macroeconomic and financial variables (Conrad & Loch, 2015). For instance, Fang, Lee, and Su (2020) try to predict the long-term stock market volatility with the GARCH-MIDAS model. In recent years, as machine learning and deep learning (Stoean, Paja, Stoean, & Sandita, 2019) have entered the financial field, related studies have gradually emerged. In detail, Milosevic (2016) and Kyriakou, Mousavi, Nielsen, and Scholz (2021) use several machine learning methods to predict long-term stock price movements. It has been proven that Random Forest is able to model financial indicators (e.g., book value, dividend yield, sales growth, etc.) with an accuracy of 76.5% (Milosevic, 2016). Moreover, Althelaya, El-Alfy, and Mohammed (2018) use bidirectional long short-term memory networks (LSTM) to evaluate and compare short- and long-term predictions of financial time series.

Long-term investing has drawn more and more attention from junior investors because of its natural advantages. Compared with short-term investments, long-term investments have lower transaction frequency, correspondingly lower transaction fees, and simpler management. In addition, it is difficult for junior investors to effectively analyze historical prices for short-term investments. However, most existing studies tend to be short-term forecasts, and only a few works carry out long-term forecasts. Therefore, there is an urgent need for high-precision long-term forecasting of stock movements for long-term investing. In China, the long-term investment decisions of investors are heavily influenced by research reports (Li & McDowell, 2011). Research reports are long, formal, and professional textual data, focusing on the value analysis of related companies. To the best of our knowledge, there are few studies (Feuerriegel & Gordon, 2018) on modeling textual data for long-term stock movement prediction. Hence, inspired by this, our work is different in that we focus on forecasting long-term stock movement by leveraging the linguistic contents of research reports. Our work falls into fundamental analysis.

2.2. Text mining for stock movement prediction

In recent years, with the significant advances in NLP, text mining technology has been widely used to process sequential textual data in forecasting tasks such as volatility and stock movements (Jaggi, Mandal, Narang, Naseem, & Khushi, 2021; Pejić Bach, Krstić, Seljan, & Turulja, 2019). Simple bag-of-words features are used to represent the financial documents in previous research (Hájek, 2018; Kumar & Ravi, 2016). It is worth noting that the assumption of independence between words, which implies a lack of grammar and word order. To make up for the inadequacy of independence assumption, recent approaches have moved towards more advanced models such as transformers (Caron & Müller, 2020) and reinforcement learning over sequential textual data for financial forecasting.

Research in recent years has examined stock movements from different sources of textual data, such as social media (Ruan, Durresi, & Alfantoukh, 2018), financial news (Zhang et al., 2018), online reviews (Li, Chen, Zhao, & Li, 2021), and firm reports (Yang et al., 2020). In detail, Xu and Cohen (2018) extract features from tweets to improve stock movement prediction. Specifically, the historical prices and tweets were jointly encoded for a stock representation. Li et al. (2020) propose the LSTM-RGCN model for financial news, which models the relationship between financial news and stock movements. Moreover, Li et al. (2021) propose a novel sentiment analysis model for stock reviews based on bidirectional encoder representation from transformers (BERT). And Yang et al. (2020) also analyze firm reports for volatility predictions based on several text mining techniques.

Despite tremendous efforts that have been made to understand the principle of stock movement prediction, most of the above studies use additional textual data as a supplement to the historical price data. Only a few attempts have been made to explore the role of textual data itself in stock movement prediction. Therefore, our work aims at finding a novel approach that can effectively predict long-term stock movements using textual data. To be specific, it is based on research reports to predict one-year stock movements. Extensive experimental results have confirmed the effectiveness of research reports in long-term stock movement prediction. In addition, we empirically show that the proposed model can significantly outperform other competing methods.

2.3. Financial studies based on research reports

In the past, most financial studies based on research reports are primarily related to investor sentiment (Chiang & Lin, 2019; Corredor, Ferrer, & Santamaria, 2019; Kim, Ryu, & Yang, 2021) or analyst sentiment (Hájek, Olej, & Myskova, 2013; Jiang et al., 2022). For instance, Song, Peng, and Huang (2020) incorporate research reports with market sentiment for stock excess return predictions. Hribar and McInnis (2012) reveal the relationship between investor sentiment and analysts' optimism. The relationship is that analysts' optimism tends to be more obvious when investor sentiment is high. Also, Yukselturk and Tucker (2015) explore the impact of analyst sentiment on UK stock recommendations and target prices. They have found a strong negative relationship between the macroeconomic and report recommendations. Moreover, Jiang et al. (2022) quantitatively measure the sentiment hidden in analyst reports of the stock market.

Existing studies based on research reports mainly focus on sentiment analysis, which requires large-scale sentiment annotations. However, it is not suitable for research reports to obtain sentiment annotations through open-domain toolkits. In addition, manual labeling is not only laborious but also expensive. To solve the above problem, we aim to construct a novel model to represent research reports using text mining techniques. Further, we attempt to make long-term stock movement predictions with the textual features of research reports. In short, our work presents a knowledge-driven approach for long-term stock movement prediction based on research reports, which avoids the problem of expensive manual annotations.

3. Problem formulation

In this work, we aim to predict the long-term movements of a target stock $s \in S$ only from the relevant research reports $r \in R$. Following Xu and Cohen (2018), we formalize stock movement prediction as a binary classification problem.

For a given set of input features X_d on the trading day d, we calculate the predicted movement Y^*_{d+T} of the closing price on the trading day $d+T$ as:

$$Y^*_{d+T} = f(X_d) \tag{1}$$

where X_d is the textual features of research reports on the trading day d, and Y^*_{d+T} is the predicted binary outcome values $\{0,1\}$ at day $d+T$. The nonlinear prediction function f maps a set of input features X_d to binary outcome Y^*_{d+T}.

For a given stock s with a closing price of p_d on the trading day d, we calculate the difference between the closing prices of the trading days d and $d+T$ as the judgment standard. Hence, the real movement Y_{d+T} of the closing price p_{d+T} on the trading day $d+T$ is defined as:

$$Y_{d+T} = \begin{cases} 1(positive), p_{d+T} > \alpha * p_d \\ 0(negative), p_{d+T} \le \alpha * p_d \end{cases} \tag{2}$$

where the hyperparameter $\alpha > 1$. We estimate the price movement binary where 1 indicates a positive increase in price volatility, and 0 means a negative increase in price volatility.

4. The SMPRR dataset

With the goal of creating a long, formal, and professional financial dataset to support research on long-term stock movement prediction, we create SMPRR that contains both research reports and historical prices. In this section, we first discuss the data construction process and its statistics. Secondly, data annotation, segmentation, and forecast periods are analyzed in detail to make better use of SMPRR. Then, the proposed dataset is compared with several popular open-source stock movement prediction datasets. The construction process of SMPRR is illustrated in Fig. 1. The dataset is available at https://github.com/zhangming-19/SMPRR. Notably, SMPRR will be continuously updated every six months for further study.

4.1. Dataset construction

We curate the dataset, SMPRR, by acquiring research reports and historical prices. The research reports between 07/23/2009 and 03/02/2022 are downloaded from the Eastmoney.com, which are publicly available on the web. Correspondingly, the historical prices of related stocks are obtained from the public finance API (https://tushare.pro/) for the last ten years. However, research reports contain not only textual data, but also lots of image data and tabular data. Unfortunately, it is difficult for machines to make better use of image data and tabular data. Therefore, we only use textual content to conduct the following research. With the help of OCR technology, the textual contents are extracted from the raw research report. It is note that several preprocessing operations are performed to improve the quality of our dataset, such as removing useless information like disclaimers and dividing research reports by themes.

4.1.1. Research reports

The dataset comprises 263,575 raw research reports. Due to the limitations of OCR, we cannot obtain the textual data of all research reports. Furthermore, the delisting of some corporations makes it impossible to gather effective data. Therefore, we have removed research reports without valid textual data or historical price data. Specifically, research reports with less than two sentences and sentences with less than five words are removed. Beyond that, to ensure the objectivity and effectiveness of data, we will no longer reduce the data easily.

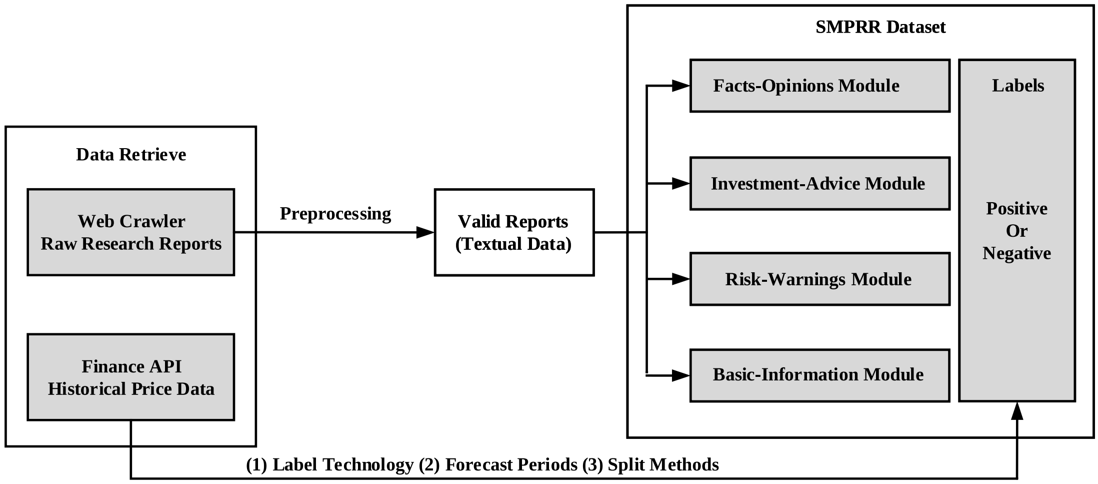

Fig. 1. The construction process of SMPRR.

Raw research reports	Processed research reports	
	Modules	Content
	FO-module	比亚迪发布2022年5月销售数据：5月份共销售汽车11.5万台，同比增长250.4%。具体来说，公司5月共销售新能源乘用车11.4万台，同比增长260.4%，环比增长8.3%。在疫情对汽车行业带来较大负面影响的背景下，这一销量水平表现相当优异… (BYD (Build-Your-Dreams) released sales data for May 2022: A total of 115,000 vehicles were sold in May, a year-on-year increase of 250.4%. In particular, the company sold a total of 114,000 new energy passenger vehicles in May, a year-on-year increase of 260.4% and a month-on-month increase of 8.3%. In the context of the relatively large negative impact of the epidemic on the automotive industry, this level of sales has performed quite well ...)
	IA-module	维持"买入"评级。我们预计公司2022-2024年净利润分别为… (The investment rating remains "Buy". We estimate that the company's profits from 2022 to 2024 will be ...)
	RW-module	补贴大幅退坡；新品推进力度不及预期。 (Subsidies have fallen sharply; The promotion of new products is less than expected.)
	BI-module	标题：月销量首破2万，高端化征程全面开启 (Title: Monthly sales exceeded 20,000 for the first time, and the high-end journey was fully launched) 股票：比亚迪 002594.SZ (Stock: BYD 002594.SZ) 行业：乘用车 (Industry: Passenger car) 评级：买入 (Rating: Buy) 评级变化：维持 (Rating change: Hold) 发布机构：安信证券 (Agency: Essence Securities) 发布作者：徐慧雄 (Author: Xu Huixiong) ...

Fig. 2. An annotation example for SMPRR dataset. We translate all Chinese texts into English for illustration.

Based on the contents of each research report, we may roughly divide it into three modules, namely Facts-Opinions, Investment-Advice, and Risk-Warnings modules. Moreover, the basic information is extracted from each report. All of the basic information is concatenated into an extra module, namely Basic-Information module. An example report is shown in Fig. 2.

As shown in Fig. 2, the processed research report is obtained from the raw research report by performing several preprocessing operations. More details of the raw research report can be found here. Additionally, the modules of the processed research report are detailed as follows:

(1) Facts-Opinions module: mainly involves a comprehensive analysis of recent activities and data of the relevant stocks, that is FO-module. It usually contains key financial news and an analysis of the company's performance. Hence, this module is typically lengthy and subjected to professional analysis.

(2) Investment-Advice module: mainly draws conclusions from the FO-module, and it can be abbreviated as IA-module. Due to the complexity and randomness of the stock market, there are a variety of ratings and recommendations. They can be roughly divided into two categories, positive and negative. More specifically, positive categories like recommendation, highly recommend, and purchase. Apart from the above positive ratings, the rest are collectively referred to as negative ratings, such as cautiously recommend, hold, and sell.

(3) Risk-Warnings module: mainly discusses potential risks that might affect the investing, that is RW-module. The RW-module is considered an important investing factor for stakeholders and the public.

(4) Basic-Information module: mainly contains the basic information, such as titles, ratings, rating changes, agencies, authors, and other basic information of the relevant stock. This module is proposed to assist the prediction of stock movements, and it can be shortened to BI-module.

To obtain high-quality databases, we manually revise all the reports by resolving duplicated or conflicted modules, dropping those without related stocks, and complementing missing basic information. As a result, we collect 161,128 high-quality research reports with 4 modules, covering 107 different industries and involving 3,671 related stocks.

Table 1
Basic statistics of SMPRR.

	Stocks	Min	Mean	Max	Tokens (millions)	Samples	Begin–End
SMPRR		60	990	4,953	159.6		
FO-module		51	757	4,561	121.0		07/23/2009
IA-module	3,671	10	151	2,028	22.6	161,128	–
RW-module		10	75	1,964	6.0		03/02/2022
BI-module		25	54	130	8.7		

4.1.2. Historical prices

The dataset contains the historical prices for each trading day from July 2005 to March 2022. In detail, the data is mainly composed of the trading date, opening price, closing price, highest price, lowest price, volume, amount, etc. The raw research reports are unlabeled. Since bias can be introduced into manually annotated datasets, and manual labeling is labor-consuming, historical prices are used for automatic annotations of research reports.

4.2. Dataset statistics and analysis

We compute the statistics of SMPRR and conduct a thorough analysis from label technology, forecast periods, and split methods. Table 1 shows an overview of the basic statistics of SMPRR.

4.2.1. Label technology

To make better use of SMPRR, label technology is discussed in detail. As we defined in the previous section, long-term stock movement prediction is transformed into a binary classification problem. Following Milosevic (2016), long-term investments are considered successful if they earn more than 10% after holding shares for a long time. Hence, we estimate the long-term stock movements binary where 1 represents an absolute increase of at least 10% in price changes, and 0 indicates a price increase of less than 10% or a price decrease. The hyperparameter α is 1.1.

As Kim, Park, and Cho (2018) pointed out, the study of capital markets needs to consider factors such as stock market closure and information leakage. In this regard, the buying price P_{buy} is determined by the closing price of the previous trading day before the release of research reports. Correspondingly, the closing price on the T-th trading day after the release of research reports is selected as the selling price P_{sell}. The label method is computed as Eq. (3):

$$L_r^{real} = \begin{cases} 1(positive), P_{sell} > 1.1 * P_{buy} \\ 0(negative), P_{sell} \leq 1.1 * P_{buy} \end{cases} \tag{3}$$

where L_r^{real} represents the real label of a research report r related to a stock. P_{buy} is the buying price of a related stock on the (d-1)-th trading day, that is $P_{buy} = p_{d-1}$. P_{sell} is the selling price of a related stock on the (d-1+T)-th trading day, that is $P_{sell} = p_{d-1+T}$.

4.2.2. Forecast periods

There has been very little research into predicting stock movements based on research reports, and we are unable to find the most reasonable forecast period in the existing studies. Intuitively, the stock movement prediction based on research reports is a long-term forecasting task, as mentioned in the IA-module "expected to rise in 6–12 months".

The purpose of research reports is stock recommendations, so investment ratings often appear in the IA-module of research reports. More specifically, the overall conclusion of the research reports on stock recommendations is summarized as investing ratings. It is note that the investment ratings represent only the subjective views of analysts, and do not bear the corresponding loss and legal responsibility. Due to the highly random and chaotic properties of the stock market, there are more than 20 different types of investment ratings in research reports, such as recommendation, strong recommendation, and cautious recommendation.

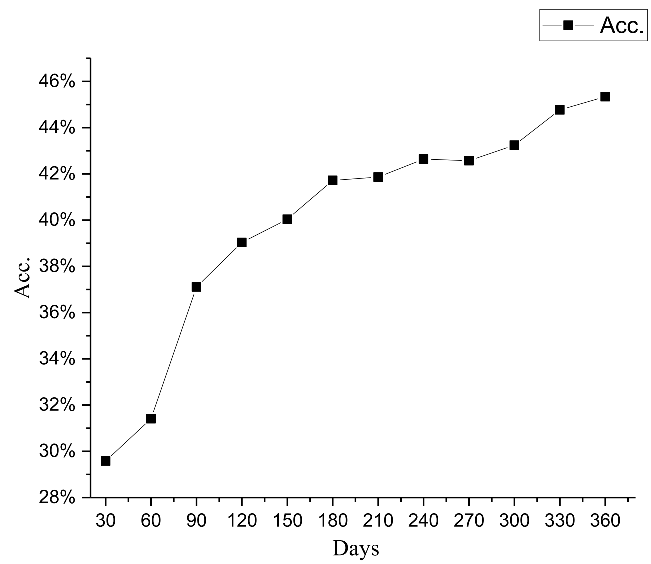

Fig. 3. The accuracy of different forecast periods of SMPRR.

Since the investment rating is the overall conclusion of a research report, it can roughly characterize the textual contents of the research report. For this reason, the best forecast period based on investment ratings is the same as that based on the textual contents of the research report. In other words, it is reasonable to calculate the accuracy of investment ratings over different periods. The best forecast period refers to the period of the highest accuracy achieved by investment ratings. Moreover, it is more convincing to beat investment ratings with the highest accuracy.

To better evaluate investment ratings, we have adopted a relatively conservative investment strategy of buying when investment ratings are positive and not otherwise. In Table 2, $L_r^{advice} = 1$, if investment ratings are positive (e.g., recommend, strongly recommend, purchase, etc.), and $L_r^{advice} = 0$, otherwise. L_r^{advice} represents the advice label of a research report r related to a stock.

In order to acquire the best forecast period of SMPRR, we first utilize Eq. (3) to label SMPRR with different periods (e.g., 30 days, 60 days, etc.) to get the real labels L_r^{real}. Secondly, we obtain the binary advice labels L_r^{advice} of investment ratings according to Table 2. Then, we use the real labels L_r^{real} and the advice labels L_r^{advice} to calculate the accuracy. Lastly, the best forecast period is the period with the highest accuracy of investment ratings. Fig. 3 shows the accuracy of investment ratings in different forecast periods.

As shown in Fig. 3, the long-term stock movement prediction based on investment ratings outperforms short-term forecasts. In detail, as the forecast period increased from 30 days to 360 days, the accuracy of investment ratings has increased from 29.6% to 45.4%. The results show that the optimal forecast period of SMPRR is 360 days, and its winning rate is 45.4%. It is note that the above winning rates are based solely on investment ratings. This works for novice investors or those who completely trust the conclusion of a research report. Obviously, it is difficult to forecast stock movements based on investment ratings alone. If interested researchers wish to employ SMPRR in future research, a forecast period of 360 days is suggested.

Table 2
The binary classification of investment ratings.

L_r^{advice}	Investment ratings
1 (positive)	Recommend, strongly recommend, purchase, overweight, outperform, etc.
0 (negative)	Cautiously recommend, neutral, sale, underweight, underperform, etc.

Table 3
Illustration of the segmentation of SMPRR.

Year	Samples	Training set		Testing set	
		Positive	Negative	Positive	Negative
2010	14,435	4,252 (34.37%)	8,121	175 (8.49%)	1,887
2011	16,853	1,642 (11.37%)	12,804	372 (15.46%)	2,035
2012	18,730	7,043 (43.87%)	9,012	1,707 (63.82%)	968
2013	16,548	7,497 (52.86%)	6,687	1,509 (63.83%)	855
2014	16,106	12,298 (89.08%)	1,508	1,905 (82.83%)	395
2015	16,018	3,558 (25.92%)	10,172	981 (42.88%)	1,307
2016	15,881	5,240 (38.49%)	8,373	664 (29.28%)	1,604
2017	20,284	4,281 (24.62%)	13,106	267 (9.22%)	2,630
2018	12,615	3,161 (29.23%)	7,652	1,112 (61.71%)	690
2019	4,921	2,785 (66.03%)	1,433	540 (76.82%)	163
Total	152,391	51,757 (39.63%)	78,868	9,232 (42.41%)	12,534

Table 4
Basic statistics of the market trends in the considered years.

Year	Market Trend	Bullish	Bearish	Turnover	Volume
2010	Bear ↓	30.67%	69.33%	304,312	25,131
2011	Bear ↓	11.95%	88.05%	237,555	21,193
2012	Bear ↓	46.72%	53.28%	164,461	18,928
2013	Bull ↑	54.42%	45.58%	229,609	26,564
2014	Bull ↑	88.18%	11.82%	375,634	42,567
2015	Bear ↓	28.34%	71.66%	1,325,588	101,702
2016	Bear ↓	37.18%	62.82%	497,865	44,884
2017	Bear ↓	22.42%	77.58%	507,770	43,799
2018	Bear ↓	33.87%	66.13%	401,965	37,235
2019	Bull ↑	67.57%	32.43%	543,844	53,792

4.2.3. Split methods

Following the standards of deep learning research, SMPRR is divided into the training set and testing set. Totally in our model and other baselines, the dataset is split chronologically in a ratio of approximately 6: 1, which is similar to Wu, Zhang, Shen, and Wang (2018). In addition, the dataset is divided by year into 10 sub-datasets from 2010 to 2019. The details of the segmentation of SMPRR are shown in Table 3.

From this table, the following findings can be drawn. Due to the high randomness and irregularity of the stock market, it is difficult to divide the dataset chronologically. The proportion of positive and negative samples in the training set and testing set is not balanced. For instance, the ratio of positive samples in the training set is much higher than in the testing set in 2010. Furthermore, the 2018 training set contains only 29.23% positive samples, compared with 61.71% in the whole testing set. However, the percentage of positive samples in the whole training set is 39.63%, which is close to 42.41% in the testing set. In this regard, the average performance of models in various market situations can be derived from numerous experiments. Due to the different market trends in the considered years, the validity of models can be demonstrated in both bear and bull markets.

Table 4 shows the basic statistics of the market trends from 2010 to 2019. In detail, bullish represents the percentage of the stock that will rise after one year, and bearish is the percentage that will fall. Moreover, turnover refers to the total transaction amount of the Shanghai Stock Exchange (SSE) in a certain year, and the unit is 100 million yuan. Volume is the total trading volume of the SSE in a certain year, and the unit is 100 million shares.

4.3. Dataset comparison

Due to the popularity of financial research, a variety of open-source datasets for predicting stock movements have been built in recent years. According to our careful research, many relevant datasets contain historical price data and textual data, such as tweets, financial news, and corporate reports. SMPRR is compared to several popular open-source datasets in Table 5. To be more specific, we first compare SMPRR with some datasets containing tweets, such as StockNet (Xu & Cohen, 2018), and CHRNN (Liu, Lin, et al., 2019). Then, SMPRR is compared to numerous financial news datasets, like LSTM-RCGN (Li et al., 2020), and EDT (Zhou, Ma, & Liu, 2021). It is worth noting that EDT is an event detection for news-based event-driven trading. Finally, SMPRR is compared to the dataset of firm annual reports, the famous Forms 10-K (Doucette & Cohen, 2015).

Table 5
A comparison between SMPRR and other existing widely-used open-source datasets.

Dataset	Samples	Stocks	Sources	Period	Language
StockNet	26,614	88	Tweets & Price	Short-term	English
CHRNN	746,286	47	Tweets & Price	Short-term	English
RGCN	18,461	–	News & Price	Short-term	English
EDT	2,265	–	Only news	Medium-term	English
Forms-10K	36,768	2,998	Only firm reports	Long-term	English
SMPRR	161,128	3,671	Research reports & Price	Long-term	Chinese

From Table 5, we observe that our dataset has some unique value compared to other datasets. The value is reflected in the following aspects: (1) SMPRR is a large-scale dataset involving lots of related stocks; (2) SMPRR is the first Chinese open-source dataset; (3) SMPRR is one of the few datasets for long-term stock movement prediction; (4) Among these datasets, the research reports is first introduced in SMPRR.

5. Methodology

We introduce the Multi-module Feature Fusion of FinBERT (MFF-FinBERT), a novel method for improving the representation of research reports, and making it better characterize the valuable contents of research reports for long-term stock movement prediction. The architecture of our model is shown in Fig. 4. It has three sub-components: Multi-module Feature Extraction, Multi-module Feature Fusion, and Stock Movement Prediction. The details of each sub-component are introduced in the following sub-sections.

5.1. Multi-module feature extraction

As mentioned in Section 2, numerous types of data (e.g., historical prices, financial news, tweets, etc.) can be used as indicators of stock price trends. As historical price data or raw textual data are insufficiently informative, feature extraction is essential to obtain meaningful representations of research reports. Based on the different modules of research reports, multi-module feature extraction is used to represent the current state of related stocks. Instead of training a model from scratch, we adopt the framework of fine-tuning a pre-trained language model on a downstream task. In particular, FinBERT is used as the feature extractor for each module.

FinBERT and RoBERTa are both variants of BERT, with a maximum input sequence length of 512. It means that BERT can only handle short texts like simple tweets, pithy news, and brief comments. For long texts, such as full news articles, or professional research reports, the maximum input sequence length limit is often exceeded. To learn more valuable information from long research reports, FinBERT is used to encode FO-module, IA-module, RW-module, and BI-module, respectively. Fig. 5 shows the architecture of Multi-module Feature Extraction based on FinBERT.

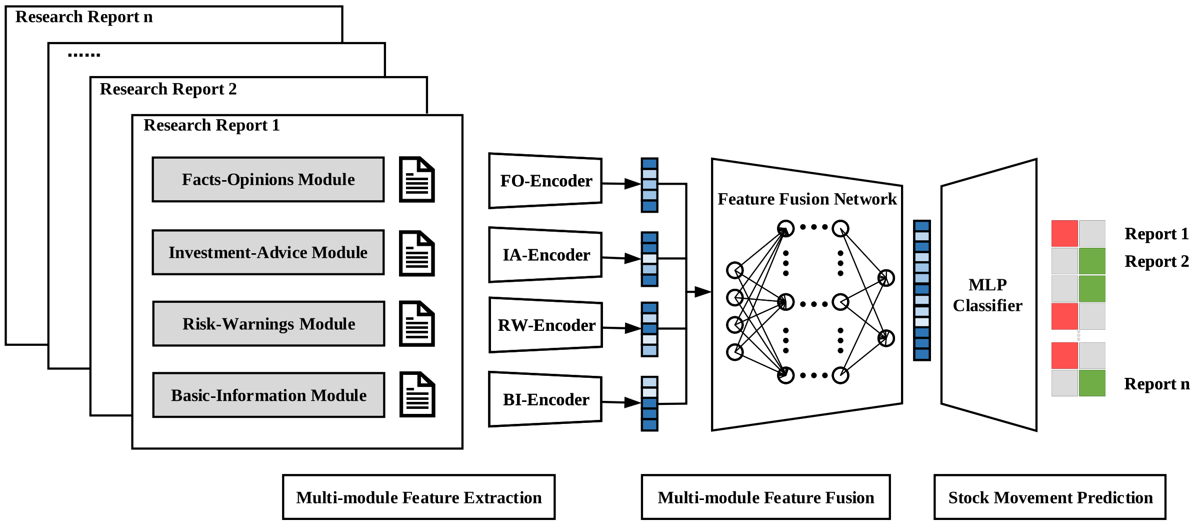

Fig. 4. General framework of MFF-FinBERT based on SMPRR.

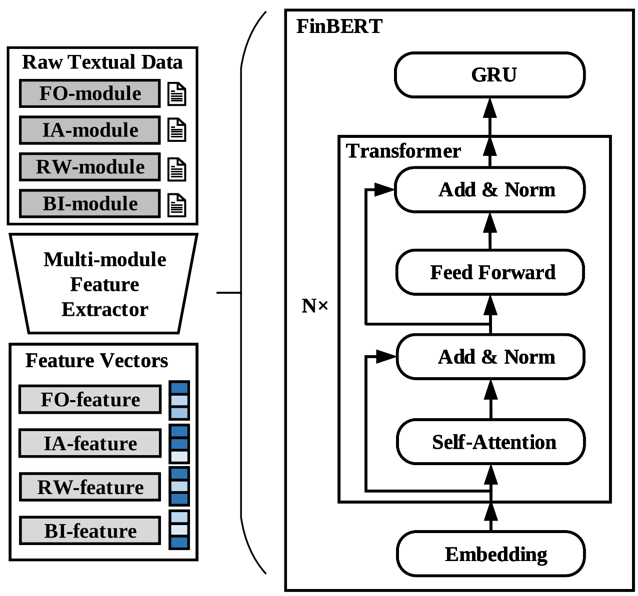

Fig. 5. An overview of Multi-module Feature Extraction based on FinBERT.

5.1.1. *Word embedding*

For each word in research reports, we need to represent it with an embedding. Following Sawhney et al. (2020), a sentence is converted into a fixed-length embedding by loading a pre-trained language model. Due to the complexity of PLMs, there are only three popular Chinese pre-trained language models, namely BERT, RoBERTa, and FinBERT. The above three PLMs are used for subsequent experiments respectively, and the influence of different PLMs is also discussed.

5.1.2. *Feature extractor*

A feature extractor is used to represent the current state of a stock based on the modules of research reports. In this study, FinBERT is used as a feature extractor to process the various modules of research reports. Therefore, there are four encoders in the proposed approach: FO-encoder, IA-encoder, RW-encoder, and BI-encoder. Specifically, FO-encoder is used to retrieve deep semantic features from the FO-module. IA-encoder encodes the IA-module. RW-encoder is used for the encoding of the RW-module. Moreover, BI-encoder obtains the encoded representations from the BI-module. All the above encoders are mainly based on FinBERT, and the basic model is transformer.

The transformer converts a sequence of symbol representations into a sequence of continuous representations. As designed in Vaswani et al. (2017), the encoder contains a stack of N = 6 identical layers. Each layer is made up of two sub-layers, one is the multi-head self-attention mechanism, and the other one is a fully connected feed-forward network. A residual connection exists around each of the two sub-layers, and layer normalization follows the two sub-layers. The primary components of transformer will not be detailed in detail, such as the multi-head self-attention mechanism, positional encoding, and scaled dot-product attention. For a more detailed explanation of the process of extracting feature vectors from raw textual data, the details could be found in Appalaraju, Jasani, Kota, Xie, and Manmatha (2021).

5.2. *Multi-module feature fusion*

Signals from different modules of research reports often provide complementary information regarding market happenings. It is unreasonable to consider merely one or two modules, and doing so may significantly skew the results. One possible approach is to directly connect the features of each module. In detail, direct concatenation

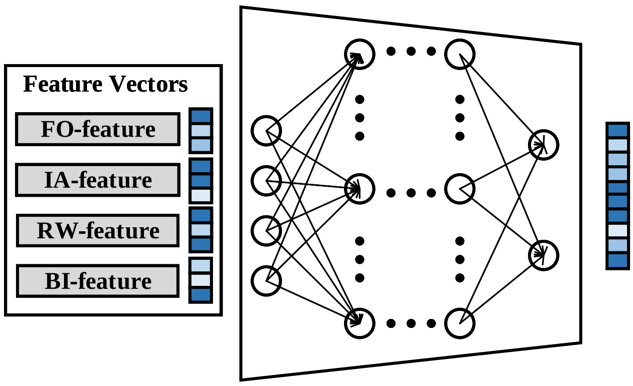

Fig. 6. The process of Multi-module Feature Fusion.

treats features from different modules equally. In this case, the voting mechanism is applied to deal with the features of various modules. If the majority of module features are positive, the vote is positive, and vice versa. However, the interdependence of modules is not adequately captured, reducing the framework's ability to learn their relevance to market movements. Towards this end, the multi-module feature fusion network is adopted to learn and weight critical module information. The process of Multi-module Feature Fusion is shown in Fig. 6.

The proposed multi-module feature fusion network processes the outputs of the multi-module feature extractor, including FO-feature, IA-feature, RW-feature, and BI-feature. The four modular features of each research report are blended to produce a joint representation. This mechanism captures the fact that modules are differently informative and have varied impacts on long-term stock movement predictions. In particular, modules with more impact information are rewarded and information from all modules is aggregated more effectively to produce better integral features.

The proposed network is designed on the basis of a BP neural network. BP neural network is a common Artificial Neural Network (ANN) algorithm, which consists of an input layer, an output layer, and one or more hidden layers. The error backpropagation technique enables BP to learn and weigh the importance of each module adaptively. Hence, the proposed multi-module feature fusion network is capable of learning features from each module and combining them to generate an aggregate representation.

5.3. Stock movement prediction

The third part is to predict stock movements based on the aggregate representation of research reports. Stock movement prediction is similar to the node classification task in previous studies. A detailed description of the problem formulation has been provided in Section 3. For the stock movement prediction task, the widely-used Multi-layer Perceptron (MLP) is used as the classification layer in this study. MLP follows the principles of the human nervous system to learn and predict. When information about the training data distribution is available, the MLP classifier is considered to be a flexible and simple classification technique. Hence, the MLP classifier is used to forecast stock movements that fall or raise.

6. Experiments

In this section, various experiments are conducted to test MFF-FinBERT based on SMPRR. First, the details of our experimental setup are provided. Then, the comparisons of the predictive performance and the computational cost between the proposed model and competing baselines are presented. Following previous studies in stock movement prediction, standard accuracy (Acc.) and Matthews Correlation Coefficient (MCC) are used as evaluation metrics. Lastly, several competing models are described in detail, including untrained models, non-pre-trained language models, and pre-trained language models.

Table 6
Parameters of the competing models.

Models	Related parameters
ARIMA	Hyperparameters p, d, q ∈ range(2).
LR	Penalty = 'l2'; tol = 0.0001; max_iter = 100.
NB	Alpha = 1.0; fit_prior = True; class_prior = None.
GRU	Epoch = 50; hidden_size = 128; num_layers = 2;
LSTM	dropout = 0.8; optimizer = adam; learning_rate = 0.01.
RoBERTa	Epoch = 3000; batch_size = 8; max_seq_length = 512;
BERT	classifier = gru; hidden_size = 512; dropout = 0.1;
FinBERT	optimizer = adam; learning_rate = 2E-5.

6.1. Dataset and training setup

We test our model and other baselines on SMPRR. In Section 4.2, the division of SMPRR and the determination of forecast period are detailed in detail. In our experiments, BERT, RoBERTa, and FinBERT are used as strong baselines. The PLMs and the word embeddings are available in Pytorch, which are used to construct the computational graph. The dimension of word embedding is 768. The maximal sequence length of our model is set to 1,024, with the excess clipped. The hidden size of the GRU layer, BP Layer, and MLP layer are set to be 512. Following Xu and Cohen (2018), we employ the input dropout rate of 0.1 to regularize latent variables. Additionally, the Adam optimizer is used to optimize the loss function with the initial learning rate of 2E-5. We fine-tune models using 3,000 epochs with a batch size of up to 8. All experiments are conducted on Tesla P100 GPU (12 GB of memory). Code and dataset could be accessed through https://github.com/zhangming-19/SMPRR.

6.2. Evaluation metrics

Following Nam and Seong (2019), two popular metrics are adopted to measure the prediction performance, including accuracy as shown in Eq. (4) and MCC as shown in Eq. (5). In particular, MCC takes the true and false positives and negatives into account to avoid bias. It is generally regarded as a balanced measure even when the data is skewed. With the confusion matrix $\begin{pmatrix} tp & fn \\ fp & tn \end{pmatrix}$, which contains the numbers of samples classified as true positive, true negative, false positive, and false negative, accuracy and MCC are calculated as follows:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{4}$$

$$MCC = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}} \tag{5}$$

6.3. Baselines

Based on SMPRR, a series of strong and representative baselines are established, including (1) Untrained models (2) Non-pre-trained language models, and (3) Pre-trained language models. We follow the same preprocessing protocols as proposed in the works and adopt their available implementations. Further, the grid search method is adopted for model parameter estimation and optimization. The parameters of the competing models are shown in Table 6. GRU and LSTM have the same parameters. In addition, the model parameters of BERT, RoBERTa and FinBERT are the same.

Untrained models (UMs): these models are not trained with past data, include:

- Research Report: a naive model that carefully forecasts stock movements based on the investment recommendations from research reports, and it can be abbreviated as RR. In this case, the prediction of stock movements is only based on the investment ratings of research reports. Hence, the input data is the investment ratings in each research report. As mentioned in Section 4.2.2, the binary forecast values of the RR model can be obtained from Table 2.
- Market Trend: a simple model that can reflect market trends, and it can be abbreviated as MT. It is well known that the market trend is a statistical result of the bullishness and bearishness of the markets. More specifically, if the real label of research report L_r^{real} is positive, the market trend is bullish and vice versa. Based on the real labels of the entire data, the market bullish percentage and bearish percentage are calculated separately. If the market bullish percentage is higher than the bearish percentage, the market trend is considered bullish. As shown in Table 4, 69.3% of the total 14435 research reports in 2010 are bearish and 30.7% are bullish, so the market trend for 2010 is bearish. It is worth noting that the MT model can be regarded as a market trend prediction model with 100% accuracy, hence it makes sense to use it as a competing baseline for comparison.
- RAND (Xu & Cohen, 2018): the simplest model that randomly guesses stock movements to be improved or declined.

Non-pre-trained language models (Non-PLMs): these approaches are not pre-trained with large-scale corpora, include:

- ARIMA (Ribeiro et al., 2021): AutoRegressive Integrated Moving Average models historical prices as a non-stationary time series. The ARIMA model is based on historical price features, including autoregression, integration, and moving average. More specifically, the autoregression is the past values of the time series. The integration represents the differential values of the time series. The moving average is the average of past values of the residual series. The ARIMA model can be formalized as:

$$\Delta^d y_t = c + \phi_1 \Delta^d y_{t-1} + \cdots + \phi_p \Delta^d y_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t \quad (6)$$

where $\Delta^d y_t$ represents the sample t of the time series y differentiated d times; p, d, and q are hyperparameters representing the model orders of the autoregression, integration and moving average, respectively; c, ϕ and θ are the model parameters; and ϵ_t represents a random white noise. The input data of ARIMA is historical prices.
- LSTM (Zolfaghari & Gholami, 2021): Long short-term memory (LSTM) is a specific type of recurrent neural network (RNN), which is suitable for processing long-term dependencies in time series. Different from the traditional RNN, LSTM regulates the flow of information through a gating mechanism. In detail, the amount of information that needs to be retained is methodically determined at each time step. Therefore, LSTM shows better performance over a long time series. Typically, an LSTM model is made up of several LSTM units, each of which contains an input gate, a forget gate, and an output gate. First, the input gate determines the information in the cell state that needs to be updated. Second, the forget gate is used to determine the information in the hidden state that should be forgotten. Finally, the output gate determines which information in the hidden state is output through the sigmoid layer. In this case, the input data of the LSTM model are historical prices and technical indicators.
- GRU (Wu et al., 2022): Gated recurrent unit is one of the popular variants of RNN. Compared with the standard LSTM, the input gate and the forget gate are combined into the update gate, and the hidden state is integrated with the cell state in GRU. Therefore, the architecture of GRU is simpler and the computational cost of GRU is less. In detail, GRU is composed of two gates, namely the update (or input) gate and the reset gate. The update (or input) gate determines the information in the cell state that should be transmitted to the next cell. The reset gate determines which information in the cell state that needs to be forgotten. Similar to LSTM, the input data of the GRU model are historical prices and technical indicators.
- Naïve Bayes (Feuerriegel & Gordon, 2018): a traditional classification model using the conditional probabilities, and it can be abbreviated as NB. For a given set of input data x_i, $i = 1, 2, \ldots, n$ with K classes, the conditional probability can be calculated as:

$$p\left(C_k \mid x_1, x_2, \ldots, x_n\right) = p\left(C_k \mid \boldsymbol{x}\right) = \frac{p\left(C_k\right) p\left(\boldsymbol{X} \mid C_k\right)}{p(\boldsymbol{x})} \quad (7)$$

where $p(\boldsymbol{x})$ is a constant value. Under the assumption that each variable is independent, the probability model can be expressed as:

$$\begin{aligned} p\left(C_k, x_1, x_2, \ldots, x_n\right) &= p\left(C_k\right) p\left(x_1 \mid C_k\right) \ldots p\left(x_n \mid C_k\right) \\ &= p\left(C_k\right) \prod_{i=1}^{n} p\left(x_i \mid C_k\right) \end{aligned} \quad (8)$$

Hence, the NB model can be defined as follow:

$$Y^* = \arg \max_{k \in \{1,\ldots,K\}} p\left(C_k\right) \prod_{i=1}^{n} p\left(x_i \mid C_k\right) \quad (9)$$

where K is 2; C_k represents the kth class; and x_i represents the vectorized features of the words in each research report.
- Logistic Regression (Cakra & Trisedya, 2015): Assuming that the data follows the Bernoulli distribution, Logistic Regression (LR) divides the data by maximizing the likelihood function and using gradient descent to solve the parameters. In detail, LR is first used to fit the input data. Then, the predicted probabilities of test data are calculated by the logistic function. Finally, the predicted probabilities are converted to binary values by the sigmoid function. Actually, we use the vectorized features of the words in each research report as the input data.

Pre-trained language models (PLMs): these methods derive common language representations from large-scale corpora and are fine-tuned for downstream tasks, which include:

- BERT (Devlin et al., 2019): Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained language model based on transformer architecture. Compared with traditional neural networks, BERT has a deeper perception of language context through bidirectional training. Bert-base-uncased is used as the textual encoder, consisting of 12 layers, 768 hidden layers, and 110M parameters. GRU is used as the classifier. The pre-built BERT model is used to fine-tune on SMPRR. As mentioned in Section 3, BERT can be formalized as:

$$Y_{d+T}^* = BERT(X_d) \quad (10)$$

where X_d is the input features of research reports on the trading day d. More specifically, the input features are composed of token embeddings, segment embeddings, and position embeddings.
- RoBERTa (Liu, Ott, et al., 2019): a Robustly optimized BERT approach (RoBERTa) is an improved version of BERT. Compared with BERT, RoBERTa adopts larger-scale training and more elaborate designs. Specifically, RoBERTa has a longer training time, a larger batch size and a large amount of training data. In addition, the training method is improved by dynamic masking and removing the next sentence prediction task. To ensure fairness, roberta-base-uncased is used as the textual encoder, and GRU is used as the classifier. It is notable that the input features are fine-tuned on SMPRR using the pre-built RoBERTa model.

Table 7
Experimental results compared with baselines. Bold shows the best performance.

	Models	Acc. ↑	MCC ↑
Untrained	RR	0.4534***	0.0275***
	RAND	0.4996***	0.0002***
	MT	0.7437*	0.0212***
Non-PLMs & Price	ARIMA	0.4763***	−0.0234***
	GRU	0.5171***	−0.0199***
	LSTM	0.5240***	−0.0308***
Non-PLMs & Reports	LR	0.7227**	0.2637***
	NB	0.7230***	0.3241***
PLMs & Reports	RoBERTa	0.7358***	0.2845***
	BERT	0.7440***	0.3317***
	FinBERT	0.7758***	0.4044***
	Ours	**0.7921**	**0.4338**

*0.1.
**0.05.
***0.01.

- FinBERT (Liu et al., 2020): a BERT-based model trained on the financial textual data, including financial news, corporate announcements, and financial encyclopedias. Same as BERT, bert-base-uncased is used as the textual encoder, and GRU is used as the classifier. Unlike BERT, the input features are fine-tuned on SMPRR using the pre-built FinBERT model.

7. Results and discussion

To understand the performance of the SOTA approach on SMPRR, we mainly discuss the experimental results and the key to predicting one-year stock movements based on research reports. In detail, the comparisons of prediction performance and computational cost are analyzed separately. Then, the importance of different modules in research reports is discussed in detail. Furthermore, the validity of the proposed components in MFF-FinBERT is investigated. Lastly, the real stock trading simulations are performed to validate the profitability of the proposed model.

7.1. Performance comparison

Details of the prediction performance of various models from 2010 to 2019 are shown in Appendix A. To test the statistical significance of the proposed model and other baselines in terms of accuracy and MCC, a group of paired t tests is performed. Table 7 shows the average prediction performance of the proposed model and other baselines on SMPRR from 2010 to 2019. In Table 7, the boldface of each column shows the best performance, and the symbol * represents the statistical significance of differences between the proposed model and other models. From this table, we have the following observations:

(1) Since RR is based solely on the investment ratings of research reports, it is concluded that long-term stock movements are difficult to predict using investment ratings. In this case, RR stands for investors who fully believe in the conclusions of research reports. There are many factors that cause the performance of RR to be worse than that of RAND, such as the impacts of sudden unexpected events.

(2) Due to the weak correlation between historical prices and long-term stock movements, it is difficult to predict one-year stock movements based on historical prices and technical indicators. As mentioned in Section 2.1, short-term forecasts are mainly technical analysis, while long-term forecasts are based on fundamental analysis. Models based on historical price and technical indicators perform significantly worse than models based on research reports, which is consistent with our expectations. In particular, the performance of ARIMA is worse than that of RAND.

(3) Introducing the representation of research reports can significantly boost the performance of predicting stock movements. First of all, compared with RR, the performance of LR has an absolute improvement of 26.93%. This demonstrates the effectiveness of using the representation of research reports for long-term stock forecasting. In addition, compared with models that make predictions based on historical prices and technical indicators, models that make predictions based on research reports improve accuracy by at least 19.87%. In other words, the performance of the report-based models is better than that of the price-based models. These experimental results confirm the validity of research reports in long-term stock movement prediction.

(4) Introducing PLMs can improve the performance of predicting stock movements, such as BERT, RoBERTa, and FinBERT. Compared with Non-PLMs, the results of PLMs are better. With an improvement of over 1.28% between PLMs and Non-PLMs, these results prove that PLMs are powerful in one-year stock movement prediction. In particular, FinBERT outperforms other baselines, reiterating that the pre-trained language model trained with financial domain texts is more effective for stock movement prediction.

(5) Introducing the multi-module feature fusion method in FinBERT can effectively improve prediction performance. Compared with FinBERT, MFF-FinBERT has an improvement of 1.63%, which reiterates the effectiveness of the multi-module feature fusion method. More details can be found in the following ablation study.

(6) In the one-year stock movement prediction task, the proposed model has achieved the best performance in terms of accuracy and MCC. It beats other competitors, including the baselines in this study and other models in previous studies. Compared to the baselines in this study, MFF-FinBERT exhibits an improvement of 1.63% and 2.94% regarding accuracy and MCC, respectively. Among these baselines, MT can be considered as a market trend prediction model with 100% accuracy. Compared with MT, the performance of the proposed model is greatly improved, including a 4.84% increase in accuracy and a 41.26% improvement in MCC. These results show that the proposed method can accurately predict market trends and effectively detect abnormal stocks. Compared with Milosevic (2016), the prediction accuracy of MFF-FinBERT is improved by 2.71%. Although not as dramatic as an improvement, these results still prove that the proposed method can better characterize research reports for long-term stock movement prediction.

(7) According to the results in Table 7, we find that there is a statistically significant difference between the proposed model and other competitive baselines, implying that our model can effectively integrate textual features from research reports to improve the performance of long-term stock movement prediction. Details of the statistical significance of differences between the proposed model and other models in terms of accuracy and MCC are shown in Appendix B.

7.2. Computational cost comparison

In real stock trading, it is very important to make accurate predictions on stock movements, grasp the exact moment of selling or buying in time, and respond quickly to sudden unexpected events. Following Dong, Wang, Luo, Wang, and Wu (2021), the computation time is adopted to measure the computational cost of competing methods. Table 8 shows the average time of predicting one target sample using various methods.

The prediction of one target sample can be decomposed into the training process and the reasoning process. Therefore, the running time of predicting one target sample is the sum of training time and reasoning time. Due to the particularity of long-term predictions based on research reports, there is no need to retrain when new research reports are available. Typically, it is enough to train PLMs monthly (or even quarterly). Therefore, the reasoning time of one target sample is a better choice than the training time in measuring the computational cost.

As shown in Table 8, the reasoning time of all models except LSTM is less than 1 min, which is feasible for low-frequency trading in the

Table 8
A comparison of computational time cost of predicting one target sample.

Models	Time cost (s)		
	Training	Reasoning	Running
RR	–	< 0.01	< 0.01
RAND	–	< 0.01	< 0.01
MT	–	< 0.01	< 0.01
ARIMA	1.68	0.01	1.69
GRU	105.51	0.86	106.37
LSTM	170.65	1.51	172.16
LR	0.21	< 0.01	0.21
NB	0.18	< 0.01	0.18
RoBERTa	2.85	0.07	2.92
BERT	1.13	0.04	1.17
FinBERT	2.71	0.08	2.79
Ours	10.08	0.27	10.35

Table 9
Experimental results of modules and their combinations. Bold shows the best performance.

Modules	Acc. ↑	MCC ↑
FO	0.7290	0.2759
IA	0.7245	0.2896
RW	0.7021	0.1130
BI	0.7492	0.3363
FO + IA	0.7367	0.3407
FO + RW	0.7292	0.3059
FO + BI	0.7769	0.3924
IA + RW	0.7267	0.2990
IA + BI	0.7486	0.3576
RW + BI	0.7593	0.3484
FO + IA + RW	0.7399	0.3457
FO + IA + BI	0.7890	0.4386
FO + RW + BI	0.7806	0.4147
IA + RW + BI	0.7789	0.3892
FO + IA + RW + BI	**0.7921**	**0.4338**

Table 10
Experimental results of ablation study. Bold shows the best performance.

Method	Settings		Metrics	
	MFE	MFF	Acc. ↑	MCC ↑
FinBERT	✗	✗	0.7758	0.4044
MFE-FinBERT	✓	✗	0.7744	0.3781
MFF-FinBERT	✓	✓	**0.7921**	**0.4338**

real stock market. Although the time cost of the proposed method is relatively larger, the performance improvement of our method makes up for this shortcoming. It is worth noting that the reasoning time of the proposed method is 0.27 s, which is more than enough for the one-year stock movement prediction task.

All experiments in this section are conducted on Tesla P100 GPU (12 GB of memory). The server has two Intel(R) Xeon(R) CPU E5-2680 v4 CPUs with 128 GB of memory. The experiments are based on the Pytorch and Tensorflow platforms. More configuration information can be found here.

7.3. Importance of different modules in research reports

To investigate the secrets of successful long-term forecasting using research reports, we conduct sufficient experiments with different modules and their combinations. The experimental results of modules and their combinations are shown in Table 9. Herein, FO, IA, RW, and BI denote the Facts-Opinions module, Investment-Advice module, Risk-Warnings module, and Basic-Information module, respectively.

From this table, we can draw the following findings. The performance of different modules and their combinations can be ranked as follows: FO + IA + RW + BI >FO + IA + BI >FO + BI >BI. We can deduce from this rank that all modules contribute to the long-term forecasting of stock movements. The importance of each module is determined in a way that minimizes performance loss by removing modules one by one. There is a slight decrease between the performance of FO + IA + RW + BI and FO + IA + BI, no more than 0.31%. Therefore, RW-module has little impact on long-term stock movement prediction. Compared with FO + IA + BI, the prediction accuracy of FO + BI is decreased by 1.21%. Lastly, there is a 2.77% gap in prediction accuracy between FO + BI and BI-module. In this way, we find the importance of various modules is listed in the following order: BI >FO >IA >RW.

In general, the RW-module is the least important for forecasting stock movements, which is consistent with human judgments. Due to the short content and serious homogeneity, it is difficult for RW-module to represent information effectively. Moreover, we can further observe that the performance improvement mainly benefits from BI-module and FO-module, while the gain from RW-module and IA-module is relatively insignificant.

Table 11
Experimental results of market performance. Bold shows the best performance.

Years	Buy-and-Hold		MFF-FinBERT		
	Net value	Yield	Net value	Yield	Trans.
2010	82.50%	−17.50%	102.73%	2.73%	109
2011	91.84%	−8.16%	138.89%	38.89%	85
2012	107.20%	7.20%	145.74%	45.74%	315
2013	102.17%	2.17%	141.96%	41.96%	298
2014	145.88%	45.88%	172.84%	72.84%	337
2015	85.69%	−14.31%	129.16%	29.16%	90
2016	111.19%	11.19%	129.98%	29.98%	201
2017	85.55%	−14.45%	105.40%	5.40%	105
2018	122.97%	22.97%	146.30%	46.30%	279
2019	117.25%	17.25%	156.33%	56.33%	185
Average	105.23%	5.23%	**136.93%**	**36.93%**	200.4

7.4. Ablation study

To explore the effectiveness of the proposed components, we further perform an incremental analysis on different settings of FinBERT. Table 10 shows the experimental results of ablation study. Compared with MFF-FinBERT, MFE-FinBERT integrates features from various modules equally, as it lacks multi-module feature fusion.

As shown in Table 10, MFE-FinBERT performs marginally worse than FinBERT, indicating that the interdependencies between modules are not properly captured. Furthermore, we can infer that the performance boost mainly benefits from multi-module feature fusion, suggesting that multi-module feature fusion can efficiently capture valuable information from each module. It has also proved the effectiveness of using MFF-FinBERT to process research reports in the stock movement prediction task.

7.5. Market performance

To further verify the profitability of the proposed model, a simulation of the profits obtained by using a trading system based on the predictions of MFF-FinBERT is performed. Taking into account the impact of market trends in different years, we conduct simulations based on SMPRR from 2010 to 2019. Specifically, a trading system based on MFF-FinBERT makes stock movement predictions for a whole year and then makes stock transactions according to a simple investment strategy. That is, buying the relevant stock when the predicted label is positive, and doing nothing when the predicted label is negative. It is note that the proposed trading strategy is compared with a buy-and-hold (BAH) investment strategy in which the investor invests with market indices. Transaction costs and taxes are not considered in this simulation.

Table A.1
Accuracy of various methods in 2010–2019. Bold shows the best performance.

Methods	Years										Average
	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	
RR	0.4180	0.3000	0.5031	0.5451	0.7659	0.3328	0.3935	0.2570	0.3638	0.6545	0.4534
RAND	0.5205	0.4968	0.5001	0.4980	0.4904	0.4977	0.4995	0.5008	0.4960	0.5137	0.5014
MT	**0.9151**	0.8455	0.6381	0.6383	0.8283	0.5712	0.7072	**0.9078**	0.6171	0.7681	0.7437
ARIMA	0.5024	0.5110	0.5013	0.4078	0.2517	0.5699	0.5851	0.8005	0.3901	0.2432	0.4763
GRU	0.7536	0.7120	0.4000	0.4184	0.2787	0.5420	0.6054	0.8212	0.3818	0.2575	0.5171
LSTM	0.8036	0.7486	0.3910	0.4044	0.2552	0.5621	0.6168	0.8398	0.3796	0.2390	0.5240
LR	0.7919	0.8455	0.6766	0.6874	0.8357	0.6180	0.7288	0.8153	0.4501	0.7781	0.7227
NB	0.6528	0.8409	0.7237	0.6667	0.8330	0.6289	0.7456	0.8053	0.5710	0.7624	0.7230
RoBERTa	0.7449	0.8671	0.7021	0.6688	0.8339	0.6106	0.7685	0.8771	0.5455	0.7397	0.7358
BERT	0.7405	0.8612	0.7529	0.6658	0.8043	0.6176	0.7650	0.8202	0.6532	0.7596	0.7440
FinBERT	0.8235	0.8675	0.7791	0.7018	0.8317	0.6477	0.7795	0.8602	0.6903	0.7767	0.7758
Ours	0.8301	**0.8729**	**0.8213**	**0.7403**	**0.8365**	**0.6481**	**0.7985**	0.8806	**0.7087**	**0.7838**	**0.7921**

Table A.2
MCC of various methods in 2010–2019. Bold shows the best performance.

Methods	Years										Average
	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	
RR	0.0408	0.0383	0.0564	0.0460	−0.0272	0.0319	0.0310	0.0310	0.0047	0.0224	0.0275
RAND	0.0060	−0.0037	−0.0005	−0.0033	−0.0116	−0.0043	−0.0001	0.0061	−0.0071	0.0208	0.0002
MT	0.0462	0.0275	0.0079	0.0084	0.0256	0.0043	0.0135	0.0370	0.0080	0.0338	0.0212
ARIMA	0.0087	0.0316	0.0041	0.0425	−0.0330	0.0521	−0.1984	−0.0421	−0.0325	−0.0674	−0.0234
GRU	0.0322	−0.0585	0.0135	0.0534	−0.0074	−0.0002	−0.1599	−0.0831	−0.0490	0.0600	−0.0199
LSTM	−0.0064	−0.0537	−0.0006	0.0470	−0.0301	0.0300	−0.1541	−0.0818	−0.0607	0.0026	−0.0308
LR	0.2825	0.0276	0.4385	0.2741	0.1858	0.2199	0.4182	0.3517	0.2010	0.2378	0.2637
NB	0.2528	0.2820	0.4784	0.3357	0.2207	0.2278	0.4494	0.3266	0.3097	**0.3573**	0.3241
RoBERTa	0.3036	0.3917	0.4808	0.3246	0.2020	0.1941	0.3769	0.0840	0.2318	0.2550	0.2845
BERT	0.3052	0.3789	0.5072	0.3261	0.2081	0.2166	0.4477	0.2961	0.3244	0.3062	0.3317
FinBERT	0.3701	0.4923	0.5450	0.3970	0.2974	0.2914	0.5197	0.4107	0.3805	0.3395	0.4044
Ours	**0.3725**	**0.5232**	**0.6225**	**0.4542**	**0.3408**	**0.2947**	**0.5207**	**0.4607**	**0.4037**	0.3452	**0.4338**

Following Wu et al. (2022), we evaluate the profitability performance of the proposed model in terms of net value to maturity and yield to maturity. The ratio of the sum of the principal and profit at maturity to the initial investment principal is known as net value to maturity. The percentage of net profit to the initial investment principal is known as yield to maturity. Therefore, net value to maturity and yield to maturity can be computed as:

$$net\_value = \frac{M_{final}}{M_{initial}} = \frac{M_{principal} + M_{profit}}{M_{principal}} = \frac{M_{exist} + M_{share}}{M_{principal}} \tag{11}$$

$$yield = \frac{M_{profit}}{M_{principal}} = \frac{M_{exist} + M_{share} - M_{principal}}{M_{principal}} \tag{12}$$

where, $M_{initial} = M_{principal}$ represents the initial investment amount of money, M_{final} is the final investment amount of money, M_{profit} denotes the amount of profit, M_{exist} is the current amount of money, and M_{share} denotes the value of all the shares held.

If the prediction of the proposed model is positive, the trading system will spend $\frac{M_{exist}}{N}$ to buy the stock at the opening price. It is worth noting that the purchase will not be executed when the minimum unit of shares (at least 100 shares) cannot be reached. In addition, during the T days that the trading system holds the stock, if the stock price meets the profit expectations $p_\theta = \frac{price_{sell} - price_{buy}}{price_{buy}}$, the system will sell immediately. For the negative forecasts, the model predicts that the underlying stock will fall, and the trading system neither buys nor sells. Herein, $M_{initial} = 1,000,000$ yuan, $N = 100$, $T = 360$, and $p_\theta = 1.1$. Table 11 shows the market performance of the proposed trading strategy and BAH on SMPRR from 2010 to 2019.

In Table 11, the boldface of each column shows the best performance, and Trans. represents the number of stock transactions. It can be seen from this table, the proposed trading strategy is superior to the buy-and-hold strategy. Generally, the net value and the yield to maturity of MFF-FinBERT strategy are higher than those of BAH strategy in different years. In bear market years (e.g., 2010, 2015, 2017, etc.), the market performance of BAH strategy is likely to be a loss, while MFF-FinBERT only shows a decrease in the yield to maturity. Compared with BAH strategy in bull market years (e.g., 2014, 2019, etc.), the performance of MFF-FinBERT strategy has been greatly improved by at least 26.69%. Also, the changes in Trans. reflect there will be more transactions during bull markets and fewer transactions during bear markets. These results demonstrate consistently better performance, which indicates the robustness of our strategy. Compared with BAH strategy, our strategy shows better market performance in terms of yield to maturity.

8. Conclusion and future work

In this paper, we explore the prediction of long-term stock movements using fused textual features, specifically predicting one-year stock movements based on Chinese research reports. To achieve this, we construct a large-scale open-source dataset named SMPRR, which consists of long, formal, and professional research reports. Based on this dataset, a series of benchmarks for long-term stock movement prediction using textual features are proposed. Furthermore, we propose MFF-FinBERT, which can effectively capture the valuable information from each module of the research reports. Extensive experiments on real market data demonstrate that MFF-FinBERT outperforms other competing baselines. In the one-year stock movement prediction task, the proposed model has achieved the best performance of 79.2%. It also signifies the effectiveness of research reports in indicating stock movements. In the simulation of trading, our strategy shows better market performance than buy-and-hold strategy in terms of yield to maturity.

In future work, we plan to conduct more exploration in both technical analysis and fundamental analysis. We are also interested in mining information contained in the historical price data to boost stock movement prediction. Meanwhile, we plan to further combine news, tweets, and other data sources to capture market dynamics better. Since model parameters have a significant impact on model performance, more efficient parameter optimization and estimation are warranted

Table B.1
The statistical significance of differences between the proposed model and other models.

Models	p-value	
	Acc.	MCC
RR	0.0002	1.84E−7
RAND	5.37E−7	3.08E−7
MT	0.0754	5.99E−7
ARIMA	0.0001	2.02E−6
GRU	0.0011	3.21E−6
LSTM	0.0026	1.68E−6
LR	0.0204	0.0018
NB	0.0035	0.0005
RoBERTa	0.0089	0.0004
BERT	0.0002	2.17E−5
FinBERT	0.0061	0.0072

in future work. Finally, we will also work on training and providing open-source models for financial problems.

CRediT authorship contribution statement

Ming Zhang: Conceptualization, Methodology, Software, Writing – original draft. **Jiahao Yang:** Software, Data curation, Investigation. **Meilin Wan:** Validation, Writing – review & editing. **Xuejun Zhang:** Visualization, Writing – review & editing. **Jun Zhou:** Funding acquisition, Project administration, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful comments and constructive guidance. This work was supported by The Youth Innovation Promotion Association of the Chinese Academy of Sciences (E1291902), Jun Zhou (2021025).

Appendix A. Details of the prediction performance of various baselines on SMPRR from 2010 to 2019

See Tables A.1 and A.2.

Appendix B. Details of the statistical significance of differences between the proposed model and other models

See Table B.1.

References

Akita, R., Yoshihara, A., Matsubara, T., & Uehara, K. (2016). Deep learning for stock prediction using numerical and textual information. In *2016 IEEE/ACIS 15th international conference on computer and information science*.

Althelaya, K. A., El-Alfy, E.-S. M., & Mohammed, S. (2018). Evaluation of bidirectional LSTM for short-and long-term stock market prediction. In *2018 9th international conference on information and communication systems*.

Appalaraju, S., Jasani, B., Kota, B. U., Xie, Y., & Manmatha, R. (2021). Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 993–1003).

Cakra, Y. E., & Trisedya, B. D. (2015). Stock price prediction using linear regression based on sentiment analysis. In *2015 international conference on advanced computer science and information systems*.

Caron, M., & Müller, O. (2020). Hardening soft information: A transformer-based approach to forecasting stock return volatility. In *2020 IEEE international conference on big data*.

Chiang, M.-T., & Lin, M.-C. (2019). Market sentiment and herding in analysts' stock recommendations. *The North American Journal of Economics and Finance*, *48*, 48–64.

Conrad, C., & Loch, K. (2015). Anticipating long-term stock market volatility. *Journal of Applied Econometrics*, *30*(7), 1090–1114.

Corredor, P., Ferrer, E., & Santamaria, R. (2019). The role of sentiment and stock characteristics in the translation of analysts' forecasts into recommendations. *The North American Journal of Economics and Finance*, *49*, 252–272.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*, *53*(9), 1375–1388.

Daudert, T. (2021). Exploiting textual and relationship information for fine-grained financial sentiment analysis. *Knowledge-Based Systems*, *230*, Article 107389. http://dx.doi.org/10.1016/j.knosys.2021.107389.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. (pp. 4171–4186). Association for Computational Linguistics.

Dong, S., Wang, J., Luo, H., Wang, H., & Wu, F. (2021). A dynamic predictor selection algorithm for predicting stock market movement. *Expert Systems with Applications*, *186*, Article 115836. http://dx.doi.org/10.1016/j.eswa.2021.115836.

Doucette, J. A., & Cohen, R. (2015). *Content of annual reports as a predictor for long term stock price movements* (pp. 416–421). AAAI Press.

Dumas, B., Kurshev, A., & Uppal, R. (2009). Equilibrium portfolio strategies in the presence of sentiment risk and excess volatility. *The Journal of Finance*, *64*(2), 579–629.

Fang, T., Lee, T.-H., & Su, Z. (2020). Predicting the long-term stock market volatility: A GARCH-MIDAS model with variable selection. *Journal of Empirical Finance*, *58*, 36–49.

Feuerriegel, S., & Gordon, J. (2018). Long-term stock index forecasting based on text mining of regulatory disclosures. *Decision Support Systems*, *112*, 88–97.

Hájek, P. (2018). Combining bag-of-words and sentiment features of annual reports to predict abnormal stock returns. *Neural Computing and Applications*, *29*(7), 343–358.

Hájek, P., Olej, V., & Myskova, R. (2013). Forecasting stock prices using sentiment information in annual reports-a neural network and support vector regression approach. *WSEAS Transactions on Business and Economics*, *10*(4), 293–305.

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, *124*, 226–251.

Hribar, P., & McInnis, J. (2012). Investor sentiment and analysts' earnings forecast errors. *Management Science*, *58*(2), 293–307.

Jaggi, M., Mandal, P., Narang, S., Naseem, U., & Khushi, M. (2021). Text mining of stocktwits data for predicting stock prices. *Applied System Innovation*, *4*(1), 13.

Jiang, X.-F., Xiong, L., Cen, T., Bai, L., Zhao, N., Zhang, J., et al. (2022). Analyst sentiment and earning forecast bias in financial markets. *Physica A: Statistical Mechanics and its Applications*, *589*, Article 126601.

Khedr, A. E., & Yaseen, N. (2017). Predicting stock market behavior using data mining technique and news sentiment analysis. *International Journal of Intelligent Systems and Applications*, *9*(7), 22.

Kim, M., Park, E. L., & Cho, S. (2018). Stock price prediction through sentiment analysis of corporate disclosures using distributed representation. *Intelligent Data Analysis*, *22*(6), 1395–1413.

Kim, K., Ryu, D., & Yang, H. (2021). Information uncertainty, investor sentiment, and analyst reports. *International Review of Financial Analysis*, *77*, Article 101835.

Kumar, B. S., & Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, *114*, 128–147.

Kyriakou, I., Mousavi, P., Nielsen, J. P., & Scholz, M. (2021). Forecasting benchmarks of long-term stock returns via machine learning. *Annals of Operations Research*, *297*(1), 221–240.

Li, W., Bao, R., Harimoto, K., Chen, D., Xu, J., & Su, Q. (2020). Modeling the stock relation with graph network for overnight stock movement prediction. In *Proceedings of the twenty-ninth international joint conference on artificial intelligence*.

Li, M., Chen, L., Zhao, J., & Li, Q. (2021). Sentiment analysis of Chinese stock reviews based on BERT model. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *51*(7), 5016–5024.

Li, W., & McDowell, E. A. (2011). Investor affect, investor status and the influence of analyst reports. *Journal of Finance and Accountancy*, *8*, 1–14.

Liu, Z., Huang, D., Huang, K., Li, Z., & Zhao, J. (2020). FinBERT: A pre-trained financial language representation model for financial text mining. In *Twenty-ninth international joint conference on artificial intelligence and seventeenth pacific rim international conference on artificial intelligence*.

Liu, J., Lin, H., Liu, X., Xu, B., Ren, Y., Diao, Y., et al. (2019). Transformer-based capsule network for stock movement prediction. In *Proceedings of the first workshop on financial technology and natural language processing* (pp. 66–73).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. CoRR abs/1907.11692, arXiv:1907.11692.

Meesad, P., & Li, J. (2014). Stock trend prediction relying on text mining and sentiment analysis with tweets. In *2014 4th world congress on information and communication technologies*.

Milosevic, N. (2016). Equity forecast: Predicting long term stock price movement using machine learning. CoRR abs/1603.00751, arXiv:1603.00751.

Nam, K., & Seong, N. (2019). Financial news-based stock movement prediction using causality analysis of influence in the Korean stock market. *Decision Support Systems*, *117*, 100–112.

Pejić Bach, M., Krstić, Ž., Seljan, S., & Turulja, L. (2019). Text mining for big data analysis in financial sector: A literature review. *Sustainability*, *11*(5), 1277.

Ribeiro, G. T., Santos, A. A. P., Mariani, V. C., & Coelho, L. d. S. (2021). Novel hybrid model based on echo state neural network applied to the prediction of stock price return volatility. *Expert Systems with Applications*, *184*, Article 115490.

Roberts, H. V. (1959). Stock-market" patterns" and financial analysis: methodological suggestions. *The Journal of Finance*, *14*(1), 1–10.

Ruan, Y., Durresi, A., & Alfantoukh, L. (2018). Using Twitter trust network for stock market analysis. *Knowledge-Based Systems*, *145*, 207–218. http://dx.doi.org/10.1016/j.knosys.2018.01.016.

Sawhney, R., Agarwal, S., Wadhwa, A., & Shah, R. (2020). Deep attentive learning for stock movement prediction from social media text and company correlations. In *Proceedings of the 2020 conference on empirical methods in natural language processing*.

Shah, D., Isah, H., & Zulkernine, F. (2019). Stock market analysis: A review and taxonomy of prediction techniques. *International Journal of Financial Studies*, *7*(2), 26.

Song, H., Peng, D., & Huang, X. (2020). Incorporating research reports and market sentiment for stock excess return prediction: A case of mainland China. *Scientific Programming*, *2020*, 8894757:1–8894757:7. http://dx.doi.org/10.1155/2020/8894757.

Souza, T. T., & Aste, T. (2019). Predicting future stock market structure by combining social and financial network information. *Physica A: Statistical Mechanics and its Applications*, *535*, Article 122343.

Stoean, C., Paja, W., Stoean, R., & Sandita, A. (2019). Deep architectures for long-term stock price prediction with a heuristic-based strategy for trading simulations. *PLoS One*, *14*(10), Article e0223593.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008.

Wu, D., Wang, X., & Wu, S. (2022). Jointly modeling transfer learning of industrial chain information and deep learning for stock prediction. *Expert Systems with Applications*, *191*, Article 116257. http://dx.doi.org/10.1016/j.eswa.2021.116257.

Wu, H., Zhang, W., Shen, W., & Wang, J. (2018). Hybrid deep sequential modeling for social text-driven stock prediction. In *Proceedings of the 27th ACM international conference on information and knowledge management* (pp. 1627–1630).

Xu, Y., & Cohen, S. B. (2018). Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th annual meeting of the association for computational linguistics (vol. 1: long papers)* (pp. 1970–1979).

Yang, Y., Zhang, K., & Fan, Y. (2020). Analyzing firm reports for volatility prediction: A knowledge-driven text-embedding approach. *INFORMS Journal on Computing*, (1), 1–19.

Yukselturk, O., & Tucker, J. (2015). The impact of analyst sentiment on UK stock recommendations and target prices. *Accounting and Business Research*, *45*(6–7), 869–904.

Yun, K. K., Yoon, S. W., & Won, D. (2021). Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process. *Expert Systems with Applications*, *186*, Article 115716.

Zhang, X., Liu, S., & Zheng, X. (2021). Stock price movement prediction based on a deep factorization machine and the attention mechanism. *Mathematics*, *9*(8), 800.

Zhang, X., Zhang, Y., Wang, S., Yao, Y., Fang, B., & Yu, P. S. (2018). Improving stock market prediction via heterogeneous information fusion. *Knowledge-Based Systems*, *143*, 236–247. http://dx.doi.org/10.1016/j.knosys.2017.12.025.

Zhao, J., Sun, N., & Cheng, W. (2020). Logistics forum based prediction on stock index using intelligent data analysis and processing of online web posts. *Journal of Ambient Intelligence and Humanized Computing*, *11*(9), 3575–3584.

Zhou, Z., Ma, L., & Liu, H. (2021). Trade the event: Corporate events detection for news-based event-driven trading. (pp. 2114–2124). Association for Computational Linguistics.

Zolfaghari, M., & Gholami, S. (2021). A hybrid approach of adaptive wavelet transform, long short-term memory and ARIMA-garch family models for the stock index prediction. *Expert Systems with Applications*, *182*, Article 115149. http://dx.doi.org/10.1016/j.eswa.2021.115149.