Neurocomputing 584 (2024) 127583

Contents lists available at ScienceDirect

Neurocomputing

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

A large-scale microblog dataset and stock movement prediction based on Supervised Contrastive Learning model

Song Yang [a,*], Daniel Tang [b]

[a] GuiZhou University of Finance and Economics, China
[b] University of Luxembourg, Luxembourg

ARTICLE INFO

Communicated by M. Gallo

Keywords:
Supervised Contrastive Learning
Stock market
A large-scale microblog dataset
Natural Language Processing

ABSTRACT

The integration of Deep Neural Networks (DNN) with Natural Language Processing (NLP) technologies has opened new avenues in financial market prediction, particularly through the utilization of textual information. This study represents a significant advancement, which offers two primary contributions to stock trend prediction: (i) the exploitation of textual data (news, comments, microblogs) using advanced DNN architectures, enhancing market information utilization; (ii) significant improvement of the accuracy of predicting the direction of stock volatility by integrating textual and neural network technologies. Meanwhile, we have crawled, filtered, and constructed a large-scale microblog dataset. This dataset includes approximately 114,992 microblog textual data from 40 Science and Technology Innovation Board (STIB) companies in China during 2021. We conducted a comprehensive analysis using various DNN techniques, including Feedback Neural Networks (FNN), Supervised Contrastive Learning (SCL), Cross Entropy (CE), and Dual Contrastive Learning (DualCL), in conjunction with bag of words models, BERT, and Roberta compilers. Our findings reveal that the SCL method, when combined with microblog data, significantly increases prediction accuracy, particularly during the COVID-19 period. Furthermore, we discovered that using a cross-stock dataset enhances the accuracy of all prediction methods, and random allocation of microblog data leads to better results than sequential allocation. Additionally, we compared the efficacy of traditional models like the CAPM, three-factor, and five-factor models against neural network-based methods. Our results suggest a notable superiority of the SCL method in increasing prediction accuracy. Finally, applying our findings to real-world trading strategies, we demonstrated the practical advantages of using the SCL method in trading, evidenced by significant improvements across all performance indicators.

1. Introduction

The prediction of stock prices has long been a subject of intense interest, which used to rely on traditional numerical data like company fundamentals and historical stock performance [1]. However, the occurrence of advanced information technology, particularly artificial intelligence and deep learning, has broadened the scope of data sources for stock market analysis. These new sources include market sentiment, policy responses, media reports, and personal social media content [2,3], all of which can significantly influence stock price trends [4].

Today, a wealth of information, ranging from market sentiment and government policies to news media reports [5] and social media content [6,7], is being harnessed to predict stock market trends. This expansion of data sources has been catalyzed by the emergence of Natural Language Processing (NLP) technologies [8,9], which have shown promising potential in deciphering the wealth of textual data available [10]. Despite the burgeoning research on stock price prediction using diverse datasets, a particular area remains underexplored: the utilization of microblog data in the Chinese market. The significance of microblogs, with their real-time updates and wide reach, cannot be overstated in a market as dynamic and interconnected as China's. However, the challenges of massive data volumes and complex collection processes have hindered comprehensive research in this area.

Recognizing this gap, our study aims to construct a significant microblog dataset from Science and Technology Innovation Board (STIB) in China for the entire year of 2021. This dataset not only enriches the sources of data for stock price prediction but also provides a unique lens through which market sentiments and trends can be analyzed. Our research employs Supervised Contrastive Learning (SCL) [11–13], a technique particularly suited for large-scale datasets, to analyze and predict stock trends based on this microblog data. This approach is

* Corresponding author.
E-mail addresses: song_yang@mail.gufe.edu.cn (S. Yang), xunzhu.tang@uni.lu (D. Tang).

https://doi.org/10.1016/j.neucom.2024.127583
Received 7 August 2023; Received in revised form 21 January 2024; Accepted 19 March 2024
Available online 23 March 2024

compared against traditional financial models such as CAPM, the three-factor, and the five-factor models, as well as other neural network methods like Feedback Neural Networks (FNN), Cross-Entropy (CE), and Dual Contrastive Learning (DualCL) [14,15]. Beyond methodological contributions, our study takes a practical approach to predicting stock market trends. We use simulated trading scenarios to test how useful our research is in real-life situations. Specifically, we look at how much profit these strategies could potentially make and how risky they are, using common measures like investment return rates and Sharpe ratios. This part of our research connects the theoretical ideas we discuss with actual stock market activities, providing useful information that can be used by both academic researchers and people who work in the finance industry.

The key contributions of this paper are as follows:

- We have constructed a comprehensive dataset encompassing microblog-related data for 40 stocks listed on China's STIB. Spanning from January 1, 2021, to December 31, 2021. This dataset aligns each stock's daily closing prices with corresponding microblog posts, resulting in an extensive collection of 114,992 data points.
- Our study leverages the SCL approach for predicting stock trends, rigorously comparing its effectiveness against other methods such as Feedback Neural Networks (FNN), Cross-Entropy (CE), and Dual Contrastive Learning (DualCL). Our analysis on both randomly sampled and time-sequenced data demonstrate the superior predictive accuracy of SCL on single stock datasets.
- We have discovered that the predictive performance of the CE, SCL, and DualCL methods is significantly improved when applied to cross-stock datasets, with DualCL showing remarkable enhancements in particular.
- The study reveals that microblog textual data is a powerful tool for enhancing the accuracy of stock trend predictions, which is especially valuable during the COVID-19 pandemic.
- We further evaluated the performance of the algorithm in real-world trading from more finance-related indicators. By conducting extensive experimental studies, we have shown that SCL achieves notable performance on these indicators.

The rest of this paper will be arranged as follows: Section 2 first summarizes the existing relevant research work on stock price trends' prediction and methods. Section 3 describes some details of the methods we use in our research. Section 4 elaborates on the planned research problem, and details the content and composition of the data set. Section 5 discusses the basic attributes of the dataset and reports the research results, and by comparing several models, it demonstrates the effectiveness of the supervised contrastive learning model. Section 6 explains the methods of dealing with the validity and limitations of the paper. Finally, Section 7 summarizes the entire paper.

2. Related work

2.1. The dataset of stock movement prediction

In the in-depth study of stock trend prediction, many scholars hold the view that there is a definite time series in stock prices. Based on this, they attempt to use historical data of stocks for prediction, which may include all kinds of structured historical data such as stock price [16], company's book-to-market value [17], stock turnover rate, company's operating income [1], and other numerical structured historical data [18,19] etc. However, with the rapid development of the Internet, we are now able to use unstructured data for stock trend prediction. Therefore, the data structure of stock trend prediction is gradually changing, expanding from traditional numerical data structure to emerging non-numerical unstructured data such as audio, video, text, images, etc. In particular, text data, due to their simplicity, recognition, processing and analysis, are favored in stock trend prediction research [20,21]. Scholars have delved into text data, including online comments [22], news [23,24], tweets [6,7] and other related text data. Some scholars even try to combine the historical numerical data of stocks with text data to jointly predict the dynamic trends of stocks [25–28]. Currently, the applied text-related data mainly include news, tweets, internal company relationships, blogs and stock-related factors. However, these materials seldom include datasets for large-scale stock microblog-related information. Therefore, constructing a dataset containing such information will undoubtedly provide significant support and contributions to future stock prediction research.

2.2. The models of stock movement prediction

There are numerous types of models for stock price prediction, often classified into models based on traditional numeric structured data and models based on unstructured data. The former primarily relies on historical numerical data for stock trend forecasting, including typical models such as the capital asset pricing model (CAPM) [16], Three-factor model [17], Five-factor model [1], and models like autoregressive conditional heteroskedasticity (ARCH) [29] and generalized autoregressive conditional heteroskedasticity (GARCH) [30]. These models primarily analyze historical numerical data, offering valuable insights into market dynamics. However, they often fall short in capturing the nuanced, real-time market sentiments that can significantly influence stock volatility.

With the rapid advancement of information technology, neural networks have gradually replaced traditional regression analysis methods, giving birth to more prediction models including artificial neural networks (ANN), Support Vector Machines (SVM) [31], multi-filtering neural networks (MFNN) [32], REGARCH-MIDAS [33], SA-DLST [34], and AE-ACG model [35]. While these models introduced advanced data processing capabilities, they still primarily relied on traditional numerical data, often overlooking the rich insights offered by unstructured data sources.

Recognizing this limitation, recent research has turned to unstructured data sources, with a particular focus on image and text data. For image data-processing, researchers proposed several models including Natural Visibility Maps (NVG) [36] and graph convolutional networks (GCN) [37]. For text data, we have the bag of words model [38], integration of attention mechanism [39], two-way Gate Recurrent Unit (GRU) networks incorporating news text noise reduction attention mechanisms based on Reinforcement Learning (RL) [40], compiler like BERT [41] and Roberta [42], hierarchical attention network based on attentive multi-view news learning (NMNL) [43], Self-supervised Contrastive Learning (SCL) [13,44], DualCL [15], and MFF-FinBERT [27]. Worth noting is that self-supervised contrastive learning offers us possibilities of avoiding the high cost of labeling large-scale datasets, dogged to be an important part of natural language processing [45,46]. Recently, Chen et al. (2022) [15] compared five benchmark text classification datasets based on NLP, DNN, and SCL technologies and concluded that DualCL could enhance classification accuracy. However, existing research has been limited by the scale and diversity of datasets used. Their reliance on only five benchmark datasets raises questions about the generalizability and stability of their findings.

Our study contributes to this field by not only utilizing a large-scale, diverse microblog dataset, but also applying and comparing a range of advanced NLP and DNN techniques, including SCL and DualCL, in the context of stock trend prediction. This approach allows us to explore the efficacy of these models on a dataset that is more representative of the actual market conditions, thereby providing a more accurate and comprehensive understanding of stock market dynamics. Moreover, by incorporating textual data from microblogs, we can capture the nuanced sentiments and trends that traditional numerical data and limited-scale unstructured data analyses miss. This comprehensive and diverse dataset approach, coupled with the application of cutting-edge NLP techniques, makes our study a significant advancement in the field of stock price prediction.

3. Approach

3.1. Problem formulation

We begin by formulating our classification problem to the context of Stock Movement Prediction. Assume a set of K distinct classes, each representing a unique stock market condition or trend. Our dataset, denoted by $D = \{(x_i, y_i)\}_{i=1}^{N}$, comprises N training samples drawn from financial markets. In this setting, x_i corresponds to an input sample with features like price, volume, or news sentiment, and y_i is the label indicating the stock market condition or trend for that sample. The index set of training samples is defined as $\mathcal{L} = \{1, 2, \ldots, N\}$, and the index set of stock market conditions or trends as $\mathcal{K} = \{1, 2, \ldots, K\}$.

This formulation sets the stage for employing contrastive learning techniques, which aim to learn powerful representations by contrasting positive and negative samples.

3.2. Supervised contrastive learning paradigm

The dominant approach to improving natural language understanding in classification tasks involves a two-stage process: pre-training and fine-tuning. In the first stage, a language model is pre-trained on an auxiliary task, which enables it to learn generic language representations. In the subsequent stage, this model is fine-tuned using task-specific labeled datasets and employs cross-entropy loss for optimization.

3.2.1. Self-supervised contrastive learning

We transition to discussing self-supervised contrastive learning, a paradigm gaining traction for its efficacy across various domains. This learning paradigm leverages unlabeled data to learn useful representations by contrasting positive (similar) and negative (dissimilar) samples. A crucial aspect of self-supervised contrastive learning is the design of the contrastive loss function, which encourages the model to minimize the distance between similar samples while maximizing the distance between dissimilar samples.

Given the N training samples $\{x_i\}_{i=1}^{N}$ along with augmented counterparts, where each sample has at least one augmented sample in the dataset, we define $j(i)$ as the index of the augmented sample derived from the ith sample. The standard contrastive loss, as per Chen et al. [15], is defined as:

$$L_{self} = \frac{1}{N} \sum_{i \in \mathcal{L}} -\log \frac{\exp(z_i \cdot z_{j(i)}/\tau)}{\sum_{\alpha \in \mathcal{A}_i} \exp(z_i \cdot z_\alpha/\tau)} \tag{1}$$

where z_i is the normalized representation of x_i, $\mathcal{A}_i := \mathcal{L} \setminus \{i\}$ is the set of indexes of the contrastive samples, the $\cdot$ symbol denotes the dot product and $\tau \in R^+$ is the temperature factor.

3.2.2. Overall objective

The aforementioned self-supervised contrastive learning scheme lacks the ability to utilize supervised signals, treating samples from the same class as positive samples and from different classes as negative samples indiscriminately [15]. To address this, we propose a supervised variant of contrastive loss [15]. The overall objective is as follows:

$$L_{sup} = \frac{1}{N} \sum_{i \in \mathcal{L}} \frac{1}{|P_i|} \sum_{p \in P_i} -\log \frac{\exp(z_p \cdot z_{j(p)}/\tau)}{\sum_{\alpha \in \mathcal{A}_i} exp(z_p \cdot z_\alpha/\tau)} \tag{2}$$

where $P_i := \{p \in \mathcal{A}_i : y_p = y_i\}$ is the set of indexes of positive samples, the $|P_i|$ is the cardinality of P_i.

3.3. Discussion on the choice of approach

The employment of a supervised contrastive learning paradigm in this work is motivated by its potential to leverage both labeled and unlabeled data in a unified framework. This hybrid approach can potentially outperform purely supervised or purely unsupervised methods, especially in scenarios where labeled data is scarce. Furthermore, the contrastive learning framework is particularly well-suited for text classification tasks, as it can effectively capture the nuanced relationships between text samples belonging to different classes.

Additionally, the incorporation of a temperature parameter in the contrastive loss functions, as expressed in Eqs. (1) and (2), provides a mechanism to control the concentration of representations, thereby offering an extra degree of flexibility in the learning process. Our choice of the supervised contrastive learning paradigm is also influenced by its demonstrated success in recent literature, where it has been shown to yield state-of-the-art performance on a variety of natural language understanding tasks. As shown in Fig. 1, a pretrained language model is first pre-trained on a general task to acquire comprehensive language representations. Following this, it is then fine-tuned with task-specific labeled datasets, where dual contrastive loss, supervised contrastive loss, and cross-entropy loss are employed simultaneously for precise model optimization.

Unlike other prevalent methods like traditional supervised learning or unsupervised learning, supervised contrastive learning offers significant advantages. Supervised learning, while powerful, often requires a substantial amount of labeled data, which can be costly and time-consuming to obtain. In situations where such data is scarce or expensive to produce, the ability of supervised contrastive learning to harness both labeled and unlabeled data becomes particularly valuable. On the other hand, unsupervised learning can leverage large amounts of unlabeled data but may fall short in achieving promising results in complex tasks, as it fails to exploit the information contained labeled data.

Moreover, supervised contrastive learning stands out for its ability to capture the nuanced differences between classes, which is often more effectively than traditional methods. By focusing on the relative comparisons between different classes, it can discern subtle distinctions and complex relationships in data, leading to more robust and generalizable models. This characteristic is particularly crucial in text classification tasks, where the intricacies of language and context play a significant role. In essence, supervised contrastive learning provides a balanced approach that is equipped with the strengths of both supervised and unsupervised learning, which makes it a superior choice for scenarios where both labeled and unlabeled data are available but labels are limited in quantity or diversity.

4. Experimental design

4.1. Research questions

In this section, we empirically evaluate our framework based on five research questions, as follows:

- (RQ1) Which training objective yields the best performance in predicting stock movement with microblog text data, the cross-entropy (CE) loss, the dual contrastive learning (DualCL) loss or the supervised contrastive learning (SCL) loss?
- (RQ2) Which method is more effective in predicting stock movement: traditional time series analysis methods using past stock history data, or neural network methods employing NLP technology with microblog text data?
- (RQ3) Does cross-stock microblog text data lead to more accurate predictions in neural network methods?
- (RQ4) What is the impact of COVID-19 on stock forecasts based on microblog text?
- (RQ5) Does SCL yield a higher rate of return under the same investment strategy compared to other neural network methods?

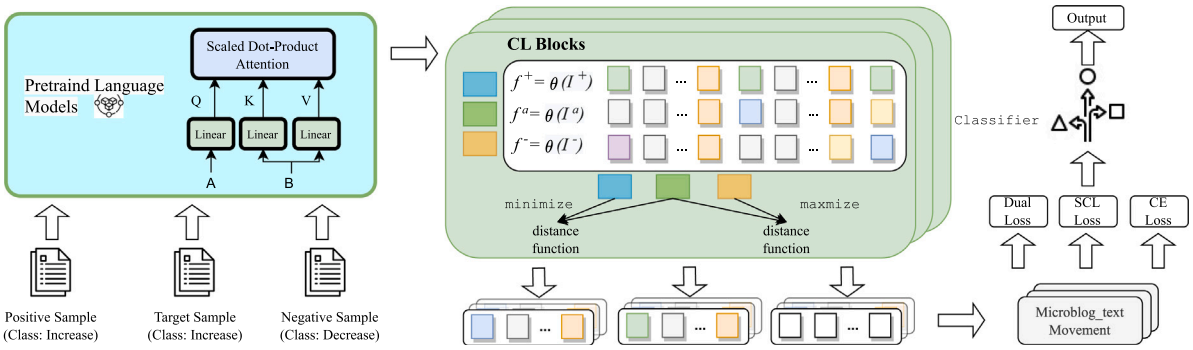


Fig. 1. The overview of our proposed objective encompasses a cross-entropy term (CE), supervised contrastive learning (SCL), and dual contrastive learning (DualCL). While we illustrate a binary classification case for simplicity, this loss framework is generally applicable to any multi-class classification setting.

Symbol	Microblog_text	Microblog_time	Closeprice	Pre_closeprice	Movement
688036	中国股市稳步上行，……	2021年01月04日 09:54	161.349	153.797	Increase
…….					
688036	周五市场高开震荡往，……	2021年12月31日 10:27	159.626	157.592	Increase

TRS : 4008*6

Extract

Microblog_text	Movement
中国股市稳步上行，……	Increase
…….	
周五市场高开震荡往，……	Increase

TRS dataset: 3970*2

Fig. 2. Overview of the raw microblog text data of TRS.

4.2. Dataset

On June 13, 2019, the Science and Innovation Board of China's stock market was officially opened. Considering that there were few listed companies on the Science and Innovation Board at the beginning and received less attention from the majority of shareholders, we mainly selected the 40 stocks from the science and innovation board of that year from January 1, 2021, to December 31, 2021 as the keywords of microblog related data search.

We obtained the stocks´ code (*Symbol*), the publishes text of microblog (*Microblogtext*), and the release time of microblog (*Microblog time*) of the 40 stocks on the Science and Technology Innovation board from the official website of microblog. Then we downloaded the daily closing price and basic informations of the 40 stocks on the Science and Technology Innovation Board stocks from the CSMAR database, and matched the release time of microblog to get the daily close price after the release time of microblog (*Closeprice*) and the daily close price before the release time of microblog (*Preclosеprice*). Furthermore, in order to analysis the (RQ2) and (RQ5), the three-factor data and five-factor data were downloaded from CSMAR database too, and Shibor radio was downloaded from the China money Network. Finally, the binary variable of movement was generated. If $Closeprice - Precloseprice > 0$, which means stock's closed price increases after the microblog text was released, then the movement is "increase"; if $Closeprice - Precloseprice < 0$, which means stock's closed price decreases after the microblog text was released, then the movement is "decrease"; if $Closeprice - Precloseprice = 0$, which means stock's closed price has no changes after the microblog text was released, then the movement is "no change". Considering that the information quality of microblog posts without stock fluctuations, which the movement is "no change", is not high, this part of data is removed. So, we obtained 114,992 stock microblog text data in the end. Take TRS stock for example in Fig. 2. Meanwhile, we have list basic information of these 40 stocks in Table 1, and draw the stock price trend chart as shown in Fig. 3.

As shown in Fig. 3, we could find the movement trend of each stock during January 1, 2021 to December 31, 2021. Some stock prices have a slightly fluctuating upward or downward trend, such as the prices of CHI, YHL and QX show a slightly fluctuating downward trend, while the prices of TE, HIU and TSE show a slightly fluctuating upward trend, and some stock prices have a large fluctuation back and forth, and no obvious trend, such as BMB, GMT and SHT. In addition, there are some stocks in the science and technology board listing time is shorter, less time data, such as HOM and ALT.

4.2.1. Microblog dataset basic information

In FNN method, we need built a words' bag by splitting the words from raw microblog text data of classify label for each stock firstly, then randomly assign 90% of data as train data and 10% of data as test data for predicting, the words number with data are listed in Table 2, and the "#UP" means the words number in increase microblog text data, the "#Down" means the words number in decrease microblog text data, and the "#Dic" means the number of words bag.

In CE, SCL and DualCL method, we computed the data length of each microblog dataset, and choose 10% of data for test data and the 90% of data for train data by randomly or chronologically. Whether dividing the data by randomly or chronologically, the number of train data and test data are unchanged. Thus, the length of dataset was listed by name of each stock. Furthermore, the microblog dataset is binary movement classification dataset about stocks of science and technology innovation board in China and each sample was labeled as "increase" or "decrease". The "no change" data, means that there was unchanged in stock price, are deleted in order to simplify the analysis. Table 2 summarizes the statistics of the dataset.

From Table 2, the "#UP" words bag almost contains about the same number of words as the "#Down" words bag, and the "#Dic" are between 1148 and 23,130 in the FNN method. Meanwhile, we can see that the volume of datasets ranged from small to large, with the smallest having only 162 samples to the largest having 5850 samples in the CE, SCL and DualCL models. And we found that there are 114,992 samples in total, including 103,483 training samples and 11,509 testing samples, and the average length of dataset is 2874.8 which is approximately 2900. Thus, 2900 is used as the dividing point between large and small samples in the following experiment.

4.3. Metric

To thoroughly evaluate the performance of the techniques and trading, we adopt the following evaluation metrics:

Table 1
Basic information about stocks.

Stocks	Symbol	ShortName	IndustryName	IndustryCode	EstablishDate	FirstListingDate
TRS	688036	Transsion	Computer, communications and other electronic equipment manufacturing	C39	2013-8-21	2019-9-30
DQE	688303	Daqo New Energy	Manufacture of electrical machinery and equipment	C38	2011-2-22	2021-7-22
CSB	688185	CanSinoBIO	Pharmaceutical industry	C27	2009-1-13	2019-3-28
BMB	688363	Bloomage Biotech	Pharmaceutical industry	C27	2000-1-3	2019-11-6
TE	688187	Times Electric	Railway, Marine, aerospace and other	C37	2005-9-26	2006-12-20
CRM	688396	CR Micro	Computer, communications and other electronic equipment manufacturing	C39	2003-1-28	2004-8-13
CM	688012	Cmsemicon	Special equipment manufacturing industry	C35	2004-5-31	2019-7-22
TSE	688599	Trinasolar	Manufacture of electrical machinery and equipment	C38	1997-12-26	2006-12-19
MT	688008	Montage Tech.	Computer, communications and other electronic equipment manufacturing	C39	2004-5-27	2013-9-26
QX	688561	Qi'an Xin	Software and information technology services	I65	2014-6-16	2020-7-22
ACM	688082	ACM Research	Special equipment manufacturing industry	C35	2005-5-17	2021-11-18
EDP	688538	Everdisplay	Computer, communications and other electronic equipment manufacturing	C39	2012-10-29	2021-05-28
VZM	688105	Vazyme	Research and experimental development	M73	2012-03-16	2021-11-15
JC	688099	Amlogic Jing Chen	Software and information technology services	I65	2003-07-11	2019-08-08
FDM	688385	FuDan Micro	Computer, communications and other electronic equipment manufacturing	C39	1998-07-10	2000-08-04
AWN	688798	Awinic	Computer, communications and other electronic equipment manufacturing	C39	2008-06-18	2021-08-16
HKT	688006	HangKe Tech.	Special equipment manufacturing industry	C35	2011-11-21	2019-07-22
VSL	688521	VeriSilicon	Software and information technology services	I65	2001-08-21	2020-08-18
HXT	688608	HengXuan Tech. Bestchnic	Computer, communications and other electronic	C39	2015-06-08	2020-12-16
ALT	688107	Anlogic	Computer, communications and other electronic equipment manufacturing	C39	2011-11-18	2021-11-12
CDB	688739	ChengDa Biotech	Pharmaceutical industry	C27	2002-06-17	2021-10-28
PNT	688063	Pylon Tech.	Manufacture of electrical machinery and equipment	C38	2009-10-28	2020-12-30
NM	688029	NanWei Medicine	Special equipment manufacturing industry	C35	2000-05-10	2019-07-22
GMT	688696	GiMi Tech.	Computer, communications and other electronic	C39	2013-11-18	2021-03-03
GDW	688390	Goodwe	Manufacture of electrical machinery and equipment	C38	2010-11-05	2020-09-04
MDC	688202	Medicilon	Research and experimental development	M73	2004-02-02	2019-11-05
CHI	688425	China Railway Construction Heavy Industry	Special equipment manufacturing industry	C35	2006-11-23	2021-06-22
HOM	688032	Hoymiles	Manufacture of electrical machinery and equipment	C38	2012-09-04	2021-12-20
OPT	688686	OPT machine Vision	Instrumentation manufacturing industry	C40	2006-03-24	2020-12-31
HFT	688200	HuaFeng Tech.	Special equipment manufacturing industry	C35	1999-09-01	2020-02-18
LRR	688499	Lyric Robot	Special equipment manufacturing industry	C35	2014-11-19	2021-07-01
CRO	688131	Chem Express	Research and experimental development	M73	2006-09-30	2021-06-08
ATS	688617	Access Point Tech. Medical	Special equipment manufacturing industry	C35	2002-06-17	2021-01-07
BPS	688368	BPS EMI	Software and information technology services	I65	2008-10-31	2019-10-14
SHT	688339	SinoHytec	Manufacture of electrical machinery and equipment	C38	2012-07-12	2020-08-10
HIU	688680	HiUV Materical Tech.	Rubber and plastic products industry	C29	2005-09-22	2021-01-22
BZ	688097	BoZhon	Special equipment manufacturing industry	C35	2006-09-22	2021-05-12
TTT	688133	TiTan Tech.	Research and experimental development	M73	2007-10-18	2020-10-30
EYL	688190	YunLu	Smelting and pressing of ferrous metals	C31	2015-12-21	2021-11-26
YHL	688575	YHLO	Pharmaceutical industry	C27	2008-09-17	2021-05-17

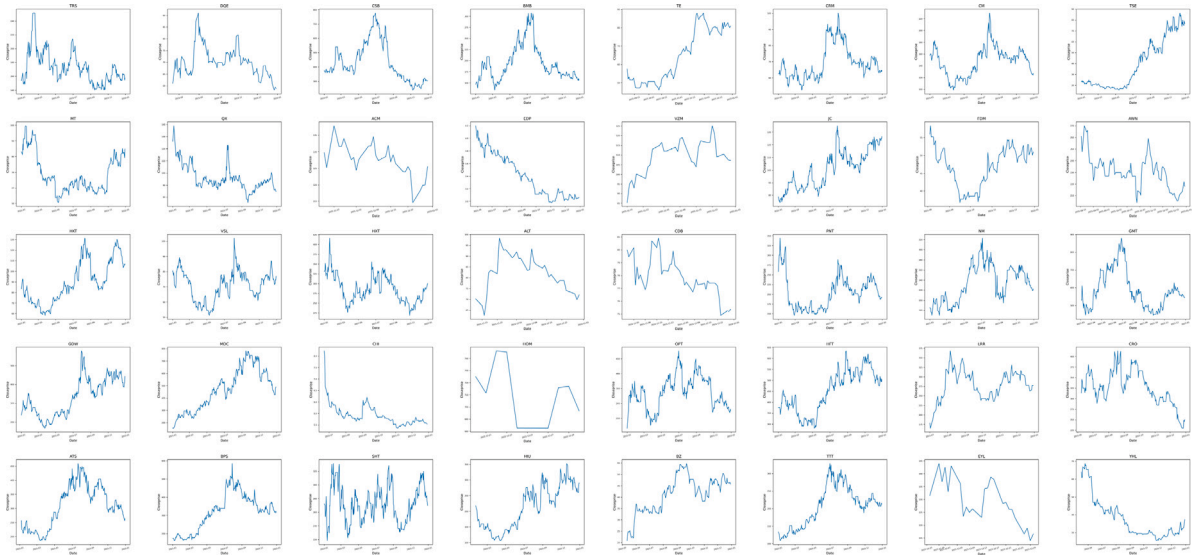


Fig. 3. Price chart of 40 stocks.

Table 2
Statistics of assigned microblog text dataset of 40 stocks.

Model	FNN			CE, SCL and DualCL (#Class = 2)						
	(#Class = 2)			Len	#Train			#Test		
Dataset	*#Up*	*#Down*	*#Dic*		*Train Len*	*#Increase*	*#Decrease*	*Test Len*	*#Increase*	*#Decrease*
TRS	102 614	103 160	12 595	3970	3573	1748	1825	397	192	205
DQ	45 896	43 431	10 241	1721	1549	775	774	172	85	87
CSB	121 245	120 476	17 945	4847	4362	2090	2272	485	218	267
BMB	144 231	161 178	23 130	5850	5265	2292	2973	585	247	338
TE	53 195	56 154	13 734	1936	1742	898	844	194	99	95
CRM	129 795	111 542	18 818	4475	4027	2114	1913	448	226	222
CM	109 893	105 636	14 961	4127	3714	1843	1871	413	204	209
TSE	117 587	102 134	15 401	4185	3766	1936	1830	419	230	189
MT	110 726	109 827	14 927	4165	3748	1936	1812	417	211	206
QX	31 024	31 847	9076	1188	1069	548	521	119	58	61
ACM	12 908	11 879	4283	487	438	237	201	49	27	22
EDP	860	3023	1148	79	71	10	61	8	1	7
VZM	14 171	56 710	5487	1940	1746	72	1674	194	6	188
JC	85 177	93 415	10 786	3516	3164	1505	1659	352	166	186
FDM	38 371	33 764	8634	1440	1296	737	559	144	79	65
AWN	33 408	37 773	10 320	1368	1231	576	655	137	61	76
HKT	112 168	94 032	13 690	3886	3497	1894	1603	389	189	200
VSL	86 488	104 783	11 776	3608	3247	1386	1861	361	162	199
HXT	87 582	101 028	11 399	3630	3267	1562	1705	363	173	190
ALT	11 368	21 735	6107	623	560	190	370	63	17	46
CDB	24 600	28 000	7787	949	854	381	473	95	37	58
PNT	118 629	121 704	23 069	4257	3831	1781	2050	426	186	240
NM	106 833	117 584	16 289	4180	3762	1638	2124	418	182	236
GMT	97 945	121 235	21 096	3871	3484	1508	1976	387	177	210
GDW	114 910	88 792	13 661	3847	3462	1970	1492	385	201	184
MDC	107 882	102 985	14 796	4188	3769	1768	2001	419	197	222
CHI	48 066	70 273	9501	2246	2021	917	1104	225	112	113
HOM	3982	4939	2318	162	145	72	73	17	10	7
OPT	102 388	110 589	10 968	3968	3571	1641	1930	397	178	219
HFT	116 377	109 091	11 990	4258	3832	2043	1789	426	213	213
LRR	45 215	47 824	9195	1793	1613	594	1019	180	72	108
CRO	63 220	79 718	10 516	2727	2454	903	1551	273	93	180
ATS	83 857	85 664	10 423	3340	3006	1395	1611	334	168	166
BPS	5357	6367	2920	239	215	115	100	24	7	17
SHT	115 415	122 386	14 796	4330	3897	2002	1895	433	219	214
HIU	93 879	86 457	12 463	3433	3089	1533	1556	344	199	145
BZ	64 892	68 923	8944	2591	2332	1082	1250	259	119	140
TTT	123 343	122 994	13 505	4569	4112	1899	2213	457	213	244
EYL	10 599	11 543	4155	442	397	191	206	45	17	28
YHL	52 337	77 959	9417	2561	2305	1059	1246	256	130	126
Sum				114 992	103 483	48 841	54 642	11 509	5381	6128

- Test accuracy: We evaluated the stock movement prediction by the Accuracy as evaluation metrics.
- Profitability: we applied the Rate of Return, Maximum Drawdown and Sharpe Ration as evaluation metrics based on our simulated trading strategy.

4.4. Baselines

We compared the performance of SCL with that of three baselines, namely CE, DualCL, and FNN. Note that SCL is a variant of contrastive learning that is implemented in a supervised learning setting. It involves learning representations by contrasting positive pairs (similar examples) against negative pairs (dissimilar examples) using labeled data. The approach enables the model to learn more nuanced and discriminative features by leveraging class label information. This results in embeddings that not only bring examples of the same class closer but also ensure that examples from different classes are distinctly separated in the feature space. The description of the baselines are as follows:

- CE: This is a standard loss function widely used in classification tasks. In our study, CE aims to maximize $z_i \cdot z_{j(i)}$ for each input x_i example by cross-entropy loss, representing the alignment between an input instance and its correct output. The objective is to increase the probability of the true class while decreasing that of the incorrect classes. This is implemented through the cross-entropy loss, as detailed in [15].
- DualCL: DualCL is an extension of the traditional contrastive learning framework. In DualCL, the goal is to maximize $z_i \cdot z_{j(i)}/\tau$ for each input x_i example by dual contrastive loss. Unlike standard contrastive learning that focuses on pulling similar examples closer and pushing dissimilar examples apart in the embedding space, DualCL introduces a dual perspective by considering additional contextual or semantic relationships. This approach is further elaborated in [15].
- FNN: The calculation of FNN is to build a dictionary set (words bag) by collecting the text data and the labels of "Up" or "Down" divided by it, and then predict the test set based on this dictionary set. And the dictionary set is sorted according to the increasing and decreasing text data, it is necessary to conduct random data sorting processing to obtain a new dictionary set. So, we initially constructed a simple three-layer feedforward neural network.

We use the AdamW [47] optimizer to finetune the pretrained BERT-base-uncased and RoBERTa-base [48] model with a 0.01 weight decay. We trained these models for 20 epochs and use a linear learn-ingrate decay 10^{-5}. We set the dropout rate to 0.1 for a layer and the train batch size to 64 and the test batch size equal to 128. And we applied the stock movement prediction by the accuracy as evaluation metrics.

Table 3
Accuracy on the test set with randomly assigned samples.

Complier	Words bag	BERT			RoBERTa		
Dataset/Model	*FNN*	*CE*	*SCL*	*DualCL*	*CE*	*SCL*	*DualCL*
TRS_r	62.96	85.64	85.64	55.67	80.86	86.15	56.93
DQE_r	66.86	93.02	93.60	54.65	90.12	79.65	55.23
CSB_r	50.00	89.90	92.37	55.26	86.39	90.93	58.35
BMB_r	63.58	94.19	93.85	72.99	62.74	92.31	57.78
TE_r	59.60	86.08	89.69	58.76	82.47	87.11	57.22
CRM_r	61.42	85.94	86.16	52.23	80.80	86.61	55.13
CM_r	59.57	86.20	90.31	53.51	72.15	86.20	58.11
TSE_r	59.24	92.84	94.99	58.23	89.50	92.84	56.32
MT_r	65.71	84.89	88.73	56.59	86.09	88.01	57.79
QX_r	76.80	98.32	96.64	56.30	98.32	94.96	70.59
ACM_r	53.70	89.80	91.84	57.14	71.43	87.76	59.18
EDP_r	66.67	87.50	87.50	87.50	87.50	87.50	87.50
VZM_r	87.13	100.00	100.00	99.48	98.97	100.00	98.97
JC_r	49.44	87.22	89.20	53.98	85.23	88.07	54.55
FDM_r	51.97	89.58	89.58	54.86	83.33	77.78	58.33
AWN_r	61.54	85.40	86.13	55.47	81.02	85.40	59.12
HKT_r	56.49	89.46	86.89	52.96	81.49	83.03	52.44
VSL_r	64.85	88.37	89.20	60.66	86.70	87.81	59.56
HXT_r	47.98	86.78	89.53	53.99	81.82	82.37	84.30
ALT_r	54.69	90.48	82.54	73.02	84.13	87.30	79.37
CDB_r	60.78	88.42	92.63	61.05	83.16	89.47	64.21
PNT_r	60.32	90.61	92.25	63.62	86.38	89.67	60.09
NM_r	60.33	88.52	89.00	56.46	81.82	85.17	56.46
GMT_r	61.24	90.70	92.76	54.52	79.33	87.60	54.52
GDW_r	61.03	60.00	91.43	56.10	82.60	87.27	53.25
MDC_r	60.00	89.26	90.69	56.32	87.35	88.54	56.32
CHI_r	54.15	86.67	87.56	50.22	76.00	88.89	53.78
HOM_r	64.71	88.24	70.59	64.71	76.47	82.35	94.12
OPT_r	54.59	85.14	86.15	55.16	81.36	81.36	55.16
HFT_r	49.07	84.98	89.44	50.47	78.64	81.46	50.47
LRR_r	64.64	83.89	85.56	60.00	83.89	85.56	60.56
CRO_r	62.95	93.04	93.04	65.93	87.55	89.01	65.93
ATS_r	64.33	88.02	85.63	49.70	82.04	61.98	51.80
BPS_r	67.74	87.50	70.83	37.50	45.83	83.33	70.83
SHT_r	63.04	87.30	91.22	56.58	80.14	87.76	61.20
HIU_r	64.64	91.86	93.02	49.71	83.72	57.85	57.27
BZ_r	64.86	81.47	86.10	54.83	81.47	83.78	55.21
TTT_r	61.00	89.28	89.06	53.39	78.56	83.81	55.58
EYL_r	62.22	77.78	68.89	62.22	66.67	66.67	64.44
YHL_r	65.63	88.28	91.02	49.22	87.89	83.98	55.86
Avg.	61.19	87.81	88.53	58.27	81.55	83.98	55.86
Len $\geq$ 2900	59.13	87.14	89.89	55.82	81.62	84.40	57.43
Len $<$ 2900	63.70	88.64	86.87	61.27	81.46	85.58	67.25
Cross-data		88.94	89.33	89.39	53.25	88.54	85.12

Note: TRS_r means the "TRS" stock dataset assigned by randomly, other stocks are similar.

5. Experimental result

5.1. The effect of supervised contrastive learning approach (RQ1 and RQ2)

5.1.1. Experimental result with randomly assigned samples

From the Table 3, in one side, compared with the FNN, CE and DualCL results, the results of SCL [13] with both BERT and RoBERTa encoders achieves better classification performance for most stocks. It also attains the best average classification performance, except in a few datasets. For instance, in the QX_r where CE with both BERT and RoBERTa encoders is used, and in the HKT_r dataset where CE with the BERT encoder is applied. Then, compared to CE, the average improvement of SCL is 0.82% and 2.98% on BERT and RoBERTa, respectively. Meanwhile, compared to DualCL, the average improvement of SCL is 51.93% and 50.34% on BERT and RoBERTa, respectively. On the other hand, we define datasets with a length greater than or equal to the mean length of all datasets (equal to 2900) as large volume samples, while the others are considered small samples. We find that in large datasets, SCL with both BERT and RoBERTa encoders achieves the best average classification performance. However, this is not the case in small datasets, indicating that SCL performs better in large datasets than in small ones.

5.1.2. Experimental result with chronologically assigned samples

To answer (RQ1) and (RQ2), considering that the data of stock trend prediction is time series prediction problem, we can calculate the training set and test set divided according to time series with historical stock price related data.

Firstly, we compute the data length of all microblog dataset respectively, and chronologically choose the former 90% of data as train data, and the latter 10% of data as test data. Two categories are still maintained, namely "Increase" and "Decrease". Therefore, the length of training set and test set remains unchanged.

Secondly, we choose the capital asset pricing model (CAPM) [16], three-factor model (3FactorM) [17] and five-factor model(5FactorM) [1] with simple linear least square estimating method for comparative analysis in traditional time series stock forecasting. These models are as following:

$$r_{it} - r_f = \alpha_i + \beta_i(r_{mt} - r_f) + e_{it} \tag{3}$$

$$r_{it} - r_f = \alpha_i + \beta_i Mkt_t + s_i SMB_t + h_i HML_t + u_{it} \tag{4}$$

$$r_{it} - r_f = \alpha_i + \beta_i Mkt_t + s_i SMB_t + h_i HML_t + r_i RMW_t + c_i CMA_t + \epsilon_{it} \quad (5)$$

where r_{it} is the return on asset i for month t, r_f is the risk-free rate (the one-year Sharpe rate), Mkt_t is the value-weight (VW) market portfolio return minus the risk-free rate, and e_{it}, u_{it}, ϵ_{it} are zero-mean residual. The remaining right-hand-side (RHS) variables are differences between the returns on diversified portfolios of small and big stocks (SMB_t), high and low B/M stocks (HML_t), stocks with robust and weak profitability (RMW_t), and stocks of low and high investment firms (conservative minus aggressive, CMA_t). If the true values of the factor exposures, α_i, β_i, s_i, h_i, r_i, and c_i, capture all differences in expected returns, the intercept α_i in (3) to (5) is indistinguishable from zero for all stocks i [49].

The CAPM model mainly analyzes the prediction of market risk premium to stock return premium. The 3FactorM mainly analyzes the prediction of market risk premium, stock market value and book market value to stock return. And the 5FactorM mainly analyzes the prediction of market risk premium, stock market value, book market value, profitability and investment pattern to stock return. All factors are weighted by market capitalization, and portfolio approach select a $2 * 2 * 2 * 2$ portfolio.

Finally, the accuracy of different stock periods, calculated using various methods, is presented in Table 4. The results indicate that, regardless of whether the BERT or RoBERTa parser is used, SCL consistently yields better and more accurate outcomes across average, large, small, and cross-sample datasets when the training and test sets are divided chronologically. Furthermore, when comparing the highest accuracy of CE, SCL, and DualCL across 40 stocks using BERT and RoBERTa analytica, it was observed that aside from 9 stocks where the accuracy was the same, BERT and RoBERTa outperformed in 16 and 15 stocks respectively, with BERT showing slightly better results than RoBERTa. From the results in Table 3, we can conclude that for the (RQ1) problem, the SCL method is most effective under both BERT and RoBERTa parsers, regardless of whether the sample set is sorted randomly or chronologically. However, the accuracy obtained in chronological order is significantly less than that obtained by random allocation. This is mainly due to the phenomenon of excessive accumulation of positive or negative information in the face of market information. As a result, when the training set and test set are allocated in chronological order, the training set is prone to lack of training samples that can better analyze the text of part of the test set, that is, when the stock price rise and fall are predicted in chronological order through the data of the microblog. The text covered by the training set is not expected to be comprehensive enough, which leads to the wrong analysis of the microblog text in the test set, which leads to the overall prediction accuracy. On the other hand, by comparing the traditional time series prediction method with Neural Networks with NLP Technology method, it can be found that the prediction effect of the latter is significantly better than that of the former. Therefore, the question (RQ2) can be answered: the neural networks of NLP technology method with microblog text data is better effective in predicting stock movement than the traditional time series analysis methods with past stock history data.

5.1.3. Visualization

It can be seen from the above research that the prediction accuracy under random allocation is significantly higher than that under time order allocation, to investigate how supervised contrastive learning improves the quality of representations, we draw the test accuracy of 20 epochs on the GMT and HIU test sets of large dataset using BERT as the compiler with randomly assigned training samples and chronologically assigned training samples. And the results of CE, SCL and DualCL methods are shown in Fig. 4.

In Fig. 4, it is observed that the test accuracy of the SCL method rapidly increases after epoch nine, significantly outperforming both the CE and DualCL methods in both the GMT and HIU datasets with randomly assigned training samples (left results). However, all the methods have no obvious advantages and disadvantages under the two large sample stocks in chronologically assigned training samples(right) results, and CE and SCL also show a downward trend when the epoch increases in GMT stocks, which indicates that the chronologically allocated sample training results have poor prediction effect.

In addition, we draw the test accuracy of 20 epochs on the QX and ALT test sets of small dataset using BERT as the encoder with randomly assigned training samples and chronologically assigned training samples. And the results are shown in Fig. 5.

In Fig. 5, the CE method performs better than both the SCL and DualCL methods in the QX and ALT datasets with randomly assigned training samples (left results). Similarly, in the two small-sample stocks with chronologically assigned training samples (right results), no method demonstrates a distinct advantage or disadvantage. Both CE and SCL exhibit a downward trend as the epoch increases in QX and ALT stocks, further indicating that training with chronologically allocated samples leads to poorer predictions.

To sum up, the following three conclusions with answers to (RQ1) and (RQ2) can be drawn. (i) SCL with both BERT and RoBERTa encoders achieves most of all stocks are better classification performance, and achieves the best classification performance in average level by randomly assigned samples or chronologically assigned samples. (ii) in case of the randomly assigned samples, accuracy of SCL method is increasing faster than the CE and DualCL methods in large datasets, but the accuracy of CE method is better in small datasets. (iii) the neural networks of NLP technology method with microblog text data are better effective in predicting stock movement than the traditional time series analysis methods with past stock history data.

5.2. The effect of cross-data (RQ3)

In order to discuss RQ3, we combined all microblog datasets by train set and test set, and got 103,483 training samples with 48,841 “increase” and 54,642 “decrease”, and 11,509 testing samples with 5381 “increase” and 6128 “decrease”, respectively. The results are showed in the bottom line of Tables 3 and 4 respectively.

It is observed that the SCL method retains the best performance with the RoBERTa encoder, whereas the DualCL method achieves an accuracy of 89.39%, the highest with the BERT encoder and better than SCL’s 88.54% accuracy, as shown in Table 3. However, in the case of chronologically ordered samples, the prediction accuracy for cross-data does not show improvement, and is even lower than the average accuracy of the CE, SCL, and DualCL models in Table 4. This suggests that the effectiveness of cross-sample predictions depends on how the training and test samples are allocated. The cross-data can effectively improve the prediction accuracy in randomly assigned samples, but it does not improve the prediction accuracy in chronological assigned sample.

Therefore, we can draw conclusion about (RQ3). Although the DualCL’s performance is not so good in individual stocks and in average level, but its performance would increase in cross stocks data. Cross-data can effectively enhance prediction accuracy in randomly assigned samples, but it does not have the same effect on chronologically assigned samples.

5.3. The impact of the COVID-19 pandemic (RQ4)

In order to analyze the impact of COVID-19 on stock trend prediction and answer (RQ4), we divided stocks into those with a high correlation with COVID-19($Highcor_{COVID-19}$) and those with a low correlation with COVID-19($Lowcor_{COVID-19}$) according to the industries mentioned in the stocks. The average forecast level of each stock is shown in Table 5.

As inferred from the table above, when using microblog data to predict stock trends, the prediction accuracy for industries closely

Table 4
Accuracy on the test set with randomly assigned samples.

Method	Traditional stock forecasting methods(OLS)			Neural networks with NLP technology					
Encoder	None			BERT			RoBERTa		
Dataset/Model	*CAPM*	tock *3FactorM*	*5FactorM*	*CE*	*SCL*	*DualCL*	*CE*	*SCL*	*DualCL*
TRS_c	46.06	43.15	40.25	56.68	58.44	52.90	57.18	56.68	56.42
DQE_c	34.58	32.71	28.04	61.63	63.95	60.47	59.30	55.23	47.67
CSB_c	43.15	33.20	32.78	53.40	55.05	56.91	56.70	54.85	54.23
BMB_c	39.42	31.54	30.71	66.84	67.69	68.21	68.03	68.03	68.21
TE_c	40.54	36.49	29.73	61.86	61.86	51.55	60.82	59.28	54.64
CRM_c	38.17	38.17	36.10	56.92	52.01	48.21	54.02	55.58	52.46
CM_c	31.95	37.76	36.10	69.98	68.52	59.32	64.89	64.89	65.86
TSE_c	39.42	40.25	34.02	54.89	55.61	51.55	56.09	51.31	48.93
MT_c	40.66	36.93	37.34	53.72	52.76	54.92	52.76	55.64	53.00
QX_c	40.25	37.76	37.34	62.18	63.03	60.50	60.50	61.34	53.78
ACM_c	48.28	44.83	41.38	85.71	93.88	87.76	100	100	87.76
EDP_c	26.21	22.76	22.07	62.50	62.50	75.00	62.50	62.50	75.00
VZM_c	40.63	31.25	37.50	100	100	100	100	100	100
JC_c	43.15	35.27	36.93	55.68	57.39	43.18	48.86	57.95	57.39
FDM_c	39.80	36.73	34.69	49.31	56.94	53.47	59.03	57.64	52.78
AWN_c	41.11	42.22	38.89	57.66	59.85	34.31	71.53	67.15	68.61
HKT_c	33.61	36.93	35.68	60.15	63.24	37.02	62.98	46.02	39.07
VSL_c	38.59	34.02	34.02	56.23	58.17	58.17	58.73	57.89	58.17
HXT_c	44.40	38.17	35.68	50.69	47.11	35.54	46.28	64.74	36.91
ALT_c	36.36	48.48	33.33	80.95	80.95	80.95	80.95	80.95	80.95
CDB_c	43.18	47.73	43.18	53.68	50.53	51.58	49.47	48.42	60.00
PNT_c	42.74	40.25	38.17	64.79	67.61	60.80	66.20	67.37	64.79
NM_c	41.08	31.95	36.10	61.24	62.92	62.92	62.92	62.92	62.92
GMT_c	38.92	35.96	36.95	60.72	60.72	60.47	60.72	60.47	60.47
GDW_c	39.00	30.29	30.29	57.66	55.58	49.61	53.51	54.03	50.13
MDC_c	36.93	36.51	36.10	59.67	59.67	53.94	60.62	62.53	60.14
CHI_c	27.91	21.71	25.58	53.33	51.56	51.11	51.11	52.89	51.11
HOM_c	42.86	42.86	28.57	64.71	58.82	76.47	58.82	64.71	70.59
OPT_c	38.17	34.02	34.44	56.93	54.66	51.64	53.65	54.91	54.41
HFT_c	40.66	38.59	40.25	54.46	53.05	53.05	53.05	53.99	55.16
LRR_c	42.62	40.98	36.07	67.22	68.33	68.33	68.33	68.33	66.67
CRO_c	35.51	34.06	34.06	70.33	70.33	70.33	70.70	70.33	70.33
ATS_c	41.77	32.91	32.49	56.29	58.08	58.38	58.38	58.08	58.08
BPS_c	41.49	39.00	41.08	50.00	58.33	62.50	41.67	62.50	54.17
SHT_c	40.66	37.76	36.10	57.04	58.89	50.35	55.43	58.43	55.43
HIU_c	38.94	35.40	34.07	56.98	60.76	49.13	55.81	59.88	46.22
BZ_c	42.68	33.12	36.94	72.59	72.59	71.81	72.20	73.75	70.27
TTT_c	41.08	36.10	33.61	63.68	61.05	61.71	62.36	62.36	62.36
EYL_c	30.43	34.78	26.09	60.00	60.00	60.00	64.44	60.00	57.78
YHL_c	38.31	30.52	31.82	40.62	41.80	34.38	34.38	34.38	34.77
Avg.	39.28	36.33	34.61	60.97	61.61	58.21	60.87	61.70	59.44
Len $\geq$ 2900	39.94	36.14	35.37	58.39	58.59	53.54	57.69	58.57	55.49
Len $<$ 2900	38.49	36.56	33.69	64.13	65.29	63.92	64.76	65.52	64.27
Cross-data				56.72	57.24	57.14	57.42	58.09	57.43

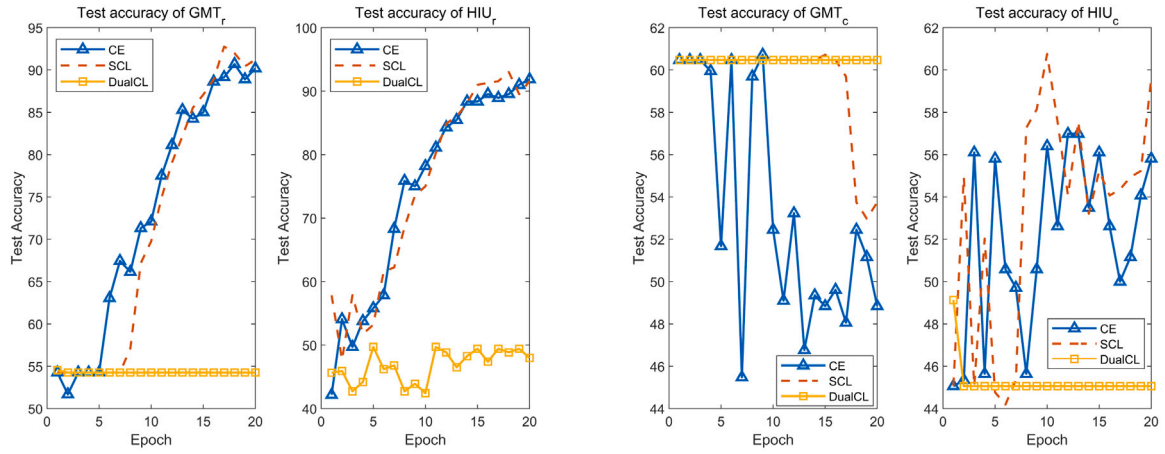
Note: TRS_c means the "TRS" stock dataset assigned by chronologically, other stocks are similar.

Table 5
Accuracy on the $Lowcor_{COVID-19}$ stocks and the $Highcor_{COVID-19}$ stocks by randomly assigned samples and chronologically assigned samples, respectively.

Method		Neural networks with NLP technology							Traditional stock return forecasting methods		
Complier		Words bag	BERT			RoBERTa			None		
Assigned	Model	*FNN*	*CE*	*SCL*	*DualCL*	*CE*	*SCL*	*DualCL*	*CAPM*	*3FactorM*	*5FactorM*
Randomly	$Lowcor_{COVID-19}$	60.51	86.88	87.46	56.79	80.92	83.73	61.28			
	$Highcor_{COVID-19}$	63.88	91.55	92.83	64.21	84.08	89.76	64.13			
Chronologically	$Lowcor_{COVID-19}$		60.33	61.19	57.23	60.40	61.47	58.36	39.16	36.63	34.53
	$Highcor_{COVID-19}$		63.53	63.27	62.13	62.78	62.61	63.76	39.78	35.11	34.97

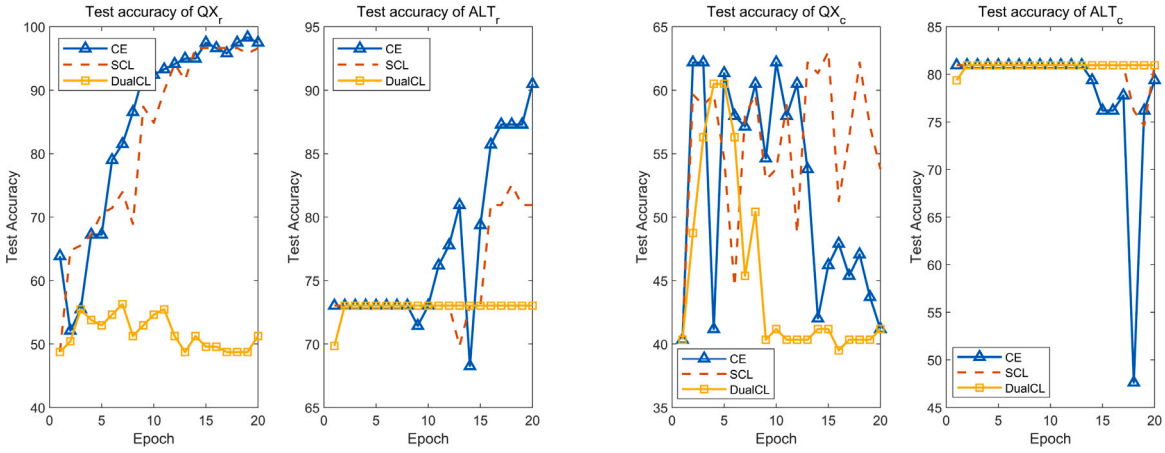
Note: The $Lowcor_{COVID-19}$ contains 8 industries, namely C29, C31, C35, C37, C38, C39, C40 and I65, while the $Highcor_{COVID-19}$ contains 2 industries, namely C27 and M73, as detailed in Table 1.

related to COVID-19, namely Research and Experimental Development (M73) and Pharmaceutical Manufacturing (C27), surpasses that of other industries, regardless of whether the data is randomized or sequential. When using traditional time series analysis methods, the degree of correlation with COVID-19 has different influences on the prediction effect of stock trend. In CAPM and 5FactorM, the prediction effect of stocks closely related to COVID-19 is better, while in 3FactorM, the opposite is true. This indicates that during the COVID-19 period, people pay more attention to stock-related microblog data, and the acquisition and comprehension of textual information exceed that of historical stock data. Therefore, when using microblog to predict stock trend, stocks in industries closely related to COVID-19 will get better



(a) Randomly assigned samples (b) Chronologically assigned samples

Fig. 4. Test accuracy on the GMT and HIU dataset in CE, SCL and DualCL using BERT with randomly assigned training samples(left) and chronologically assigned training samples (right).



(a) Randomly assigned samples (b) Chronologically assigned samples

Fig. 5. Test accuracy on the QX and ALT dataset in CE, SCL and DualCL using BERT with randomly assigned training samples(left) and chronologically assigned training sample (right).

prediction effect. However, with traditional time series analysis methods, the effectiveness is inconsistent regarding the stocks' correlation with COVID-19.

5.4. The return of stocks (RQ5)

5.4.1. Simulated trading strategy

We refer to the investment strategies in this article [23] and specifically divide investment operations into three categories: short, long, and preserve. Given that some stocks in the Chinese stock market are subject to short selling, we assume that a short selling policy applies to all stocks for our analysis. When the stock is predicted to fall in the next trading day, the "short" operation is performed in current trading day. When the stock is predicted to rise in the next trading day, the "long" operation is performed in current trading day. Otherwise, when there is no change in the next trading day or there is no microblog information about the stock in current trading day, the "preserve" operation is performed. And for different trading operations, different yield calculation methods are adopted. When the "short" operation is executed, the log of daily return rate is defined as $r_t = [ln(P_{sell}) - ln(P_{buy})] * 100\%$. When the "long" operation is executed, the log of daily return rate is defined as $r_t = [ln(P_{buy}) - ln(P_{sell})] * 100\%$. And when the "preserve" operation is executed, the log of daily return rate is $r_t = 0$. So, the monthly rate of return is $e^{sum(r_t)} - 1$. P is the stock close price, and $ln(P)$ is the log function of P.

This strategy is based on the assumption that the stock price will surge or plummet in response to the release of microblog information. The trading operations will be triggered according to the result predicted by our model. Specifically, if the average predicted result from microblog texts is greater than or equal to 0 but less than a predefined threshold (indicating a "downtrend" signal), a "short" operation is triggered. Conversely, if the average result is greater than or equal to this threshold and less than 1.5 (indicating an "uptrend" signal), a "long" action is initiated. With different dataset's predicted results, the threshold would be different. See Table 6.

Table 6
The comparison of profitability test on Maximum Drawdown (%), Daily Rate of Return (%), and Sharpe Ratio Rate with strong baselines, GMT, HIU, QX and ALT of BERT model from 11/1/2021 to 12/31/2021.

Dataset	Microblog period	Train interval	Test interval	Method	Monthly Rate of Return (%)	Maximum Drawdown (%)	Sharpe Ratio
GMT	2021/3/3-2021/12/31	[1,3433]	[3434,3871]	SCL	4.38	10.59	1.92
				CE	4.13	10.59	1.59
				DualCL	3.61	10.59	0.94
HIU	2021/1/25-2021/12/31	[1,3046]	[3047,3463]	SCL	9.94	16.71	3.65
				CE	2.33	16.71	−0.27
				DualCL	0.81	16.71	−1.04
QX	2021/1/3-2021/12/31	[1,1131]	[1132,1228]	SCL	4.04	15.79	1.81
				CE	3.38	15.79	0.77
				DualCL	−1.42	15.79	−6.42
ALT	2021/11/15-2021/12/31	[1,217]	[218,639]	SCL	20.18	25.03	11.71
				CE	21.37	25.03	12.73
				DualCL	1.46	25.03	−0.83

5.4.2. Profit metrics

For simulated trading, we applied the Daily Rate of Return, Maximum Drawdown and Sharpe Ratio as evaluation metrics based on our simulated trading strategy.

- Monthly Rate of Return reflects the monthly return level of the stock, and $MonthlyRateofReturn = e^{sum(r_t)} - 1$. $e^{(\cdot)}$ means to the exponential function, and $sum(\cdot)$ means to the summation function.
- Maximum Drawdown is a risk measure of the degree to which an asset holds its value, and $MaximumDrawdown = Max(P_x - P_y)/P_x * 100\%, with y > x$.
- Sharpe Ratio reflects its nature of balancing return and risk of a strategy, and $SharpeRatio = (RateofReturn - Risk - freeInterest Rate)/StandardDeviationofReturn$, which Risk-free Interest Rate equals to 0.0287 that is the average 1Y Shibor rate in 2021 year.

5.4.3. Simulation result

In order to analysis the performance of return by different methods in real stocks, we compare the profit of test with GMT, HIU, QX, and ALT. The length of Train and Test data information and simulation results are summarized in Table 6.

In this section, we discuss the possible profitability of the proposed strategy in real-world trading. We use our trading strategy to conduct trading simulation (backtesting) on stock data from November 2021 to December 2021 using the stock movement prediction result of our model trained from January 2021 to October 2021 as mentioned in Section 5.2. In Table 6, we demonstrate that the ALT datasets, when processed with SCL and CE methods, achieve an impressive monthly rate of return of 20.18% and 21.37%, respectively, surpassing the others. Meanwhile, the average monthly return rate of these four datasets with three methods are 9.635%, 7.8025%, and 1.115%, respectively, with SCL being the highest. Furthermore, SCL exhibits superior profit performance, a lower Maximum Drawdown, and a higher Sharpe Ratio in large datasets; however, these advantages are not as pronounced in smaller datasets.

6. Discussion

6.1. Threat to validation

This paper constructs a large-scale stock microblog dataset for forecasting stock trends. This dataset serves as a reliable basis for further research in stock prediction using natural language processing. Compared to the study by [15], our larger dataset yields distinct results when employing BERT and RoBERTa language models. We posit that SCL offers a superior predictive effect.

6.2. Limitations

The research presented in this paper has some limitations. It focuses only on predicting stock trends based on stock microblog data, overlooking the significant correlation between stocks. Moreover, in China, stock trend fluctuations are also influenced by other types of information such as news, policy documents, and newspaper reports. Relying solely on stock microblog data for predictions is somewhat simplistic. Therefore, integrating these factors represents a valuable direction for future research enhancement. Additionally, our analysis of stock prediction accuracy across different methods involves comparing text data of varying sample sizes. We find that both large and small samples affect the accuracy of single stock predictions. Notably, there are significant differences in the predictive outcomes between cross-sample and single-stock sample datasets.

7. Conclusion and future work

7.1. Conclusion

In this paper, we build a platform with microblog dataset and study the problem of NLP-based stock movement prediction. We have drawn multiple conclusions as follows:

- SCL with both BERT and RoBERTa encoders generally achieves better classification performance for most stocks, showing the best average performance in both randomly and chronologically assigned samples. However, SCL performs better in large datasets than in small ones when samples are randomly assigned.
- Neural network methods using NLP technology with microblog text data are more effective in predicting stock movement than traditional time series analysis methods using historical stock data.
- The cross stocks data would improve the accuracy of CE, SCL and DualCL methods, especially in DualCL, in randomly assigned samples, but it does not improve the prediction accuracy in chronological assigned samples.
- In the period of COVID-19, people pay more attention to stock-related microblog data, and the acquisition and recognition of textual information are higher than that of historical stock data. Therefore, when using microblog to predict stock trend, stocks in industries closely related to COVID-19 will get better prediction effect. However, when using traditional time series analysis methods, the effectiveness is inconsistent regardless of the stock's relation to COVID-19.
- The SCL reflects better profit performance, smaller Maximum Drawdown and larger Sharpe Ratio in the large datasets, but it is not necessarily true in the small datasets.

7.2. Future work

In this paper, through a web crawl blog data and sorting and build a larger microblog dataset, then stocks movement are discussed by the natural language technology with neural networks methods of FNN, CE, SCL and DualCL, and traditional stocks return time series forecast models of CAPM, Three-factor Model and Five-factor Model. However, this article has not addressed two points. The first is the potential significant relationship between different textual data in microblogs. The second is the possible correlation between different stocks. These two points have not considered in this paper and can be used as the direction of future research.

CRediT authorship contribution statement

Song Yang: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Daniel Tang:** Methodology, Project administration, Supervision, Validation, Visualization, Writing – review & editing.

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.

Acknowledgements

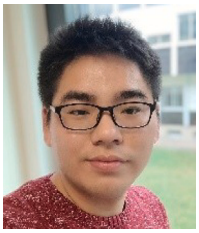
The authors thank the editors and reviewers for their comments, which led to the improvement of this paper. This work is supported by the Guizhou Province Philosophy and Social Science Foundation of China (Grant No. 22GZQN03).

References

[1] E.F. Fama, K.R. French, A five-factor asset pricing model, J. Financ. Econ. 116 (2013) 1–22.
[2] Rui Cheng, Qing Li, Modeling the momentum spillover effect for stock prediction via attribute-driven graph attention networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 55–62.
[3] Murat Guven, Basak Cetinguc, Bulent Guloglu, Fethi Calisir, The effects of daily growth in COVID-19 deaths, cases, and governments' response policies on stock markets of emerging economies, Res. Int. Bus. Finance 61 (2022) 101659.
[4] Liheng Zhang, Charu Aggarwal, Guo-Jun Qi, Stock price prediction via discovering multifrequency trading patterns, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 2141–2149.
[5] Yelin Li, Hui Bu, Jiahong Li, Junjie Wu, The role of text-extracted investor sentiment in Chinese stock price prediction with the enhancement of deep learning, Int. J. Forecast. 36 (4) (2020) 1541–1562.
[6] Bing Li, Keith C.C. Chan, Carol Ou, Sun Ruifeng, Discovering public sentiment in social media for predicting stock movement of publicly listed companies, Inf. Syst. 69 (2017) 81–92.
[7] S. Mehtab, J. Sen, A robust predictive model for stock price prediction using deep learning and natural language processin, 2019, arXiv preprint arXiv:1912.07700.
[8] B. Alshemali, J. Kalita, Improving the reliability of deep neural networks in NLP: A review, Knowl.-Based Syst. 191 (2020) 105210.
[9] B. Min, H. Ross, E. Sulem, et al., Recent advances in natural language processing via large pre-trained language models: A survey, ACM Comput. Surv. 56 (2) (2023) 1–40.
[10] F.Z. Xing, E. Cambria, R.E. Welsch, Natural language based financial forecasting: a survey, Artif. Intell. Rev. 50 (2018) 49–73.
[11] P. Khosla, P. Teterwak, C. Wang, et al., Supervised contrastive learning, Adv. Neural Inf. Process. Syst. 33 (2020) 18661–18673.
[12] S. Wan, S. Pan, J. Yang, et al., Contrastive and generative graph convolutional networks for graph-based semi-supervised learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 10049–10057, (11).
[13] Beliz Gunel, Jingfei Du, Alexis Conneau, Veselin Stoyanov, Supervised contrastive learning for pre-trained language model fine-tuning, in: International Conference on Learning Representations, ICLR, 2021.
[14] M. Sailer, E. Bauer, R. Hofmann, et al., Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning, Learn. Instr. 83 (2023) 101620.
[15] Q. Chen, R. Zhang, Y. Zheng, et al., Dual contrastive learning: Text classification via label-aware data augmentation, 2022, arXiv preprint arXiv:2201.08702.
[16] W.F. Sharpe, Capital asset prices: A theory of market equilibrium under conditions of risk, J. Finance 19 (3) (1964) 425–442.
[17] E.F. Fama, K.R. French, Common risk factors in the returns on stocks and bonds, J. Financ. Econ. 33 (1993) 3–56.
[18] Jigar Patel, Sahil Shah, Priyank Thakkar, K. Kotecha, Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques, Expert Syst. Appl. 42 (1) (2015) 259–268.
[19] Bowen Pang, Wei Wei, Xing Li, Xiangnan Feng, Chao Li, A representation-learning-based approach to predict stock price trend via dynamic spatiotemporal feature embedding, Eng. Appl. Artif. Intell. 126 (Part A) (2023) 106849.
[20] O. Mane, Stock market prediction using natural language processing–A survey, 2022, arXiv preprint arXiv:2208.13564.
[21] O. Bustos, A. Pomares-Quimbaya, Stock market movement forecast: A Systematic review, Expert Syst. Appl. 156 (2020) 113464.
[22] Junran Wu, Ke Xu, Jichang Zhao, Predicting long-term returns of individual stocks with online reviews, Neurocomputing 417 (2020) 406–418.
[23] J. Zou, H. Cao, L. Liu, et al., Astock: A new dataset and automated stock trading based on stock-specific news analyzing model, 2022, arXiv preprint arXiv:2206.06606.
[24] Heyan Huang, Xiao Liu, Yue Zhang, Chong Feng, News-driven stock prediction via noisy equity state representation, Neurocomputing 470 (2022) 66–75.
[25] Xiaodong Li, Xiaodi Huang, Xiaotie Deng, Shanfeng Zhu, Enhancing quantitative intra-day stock return prediction by integrating both market news and stock prices information, Neurocomputing 142 (2014) 228–238.
[26] R. Sawhney, S. Agarwal, A. Wadhwa, et al., 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, EMNLP, 2020, pp. 8415–8426.
[27] Ming Zhang, Jiahao Yang, Meilin Wan, Xuejun Zhang, Jun Zhou, Predicting long-term stock movements with fused textual features of Chinese research reports, Expert Syst. Appl. 210 (2022) 118312.
[28] Yu Ma, Rui Mao, Qika Lin, Peng Wu, Erik Cambria, Multi-source aggregated classification for stock price movement prediction, Inf. Fusion 91 (2023) 515–528.
[29] R.F. Engle, Autoregressive conditional heteroskedasticity with estimates of the variances of U.K. inflation, Econometrica 50 (1982) 987–1008.
[30] Tim Bollerslev, Generalized autoregressive conditional heteroskedasticity, J. Econometrics 21 (1986) 307–328.
[31] Yakup Kara, Melek Acar Boyacioglu, Ömer Kaan Baykan, Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange, Expert Syst. Appl. 38 (5) (2011) 5311–5319.
[32] Wen Long, Zhichen Lu, Lingxiao Cui, Deep learning-based feature engineering for stock price movement prediction, Knowl.-Based Syst. 164 (2019) 163–173.
[33] Lu Wang, Chenchen Zhao, Chao Liang, Song Jiu, Predicting the volatility of China's new energy stock market: Deep insight from the realized EGARCH-MIDAS model, Finance Res. Lett. 48 (2022) 102981.
[34] Yanli Zhao, Guang Yang, Deep Learning-based Integrated Framework for stock price movement prediction, Appl. Soft Comput. 133 (2023) 109921.
[35] Shicheng Li, Xiaoyong Huang, Zhonghou Cheng, Wei Zou, Yugen Yi, AE-ACG: A novel deep learning-based method for stock price movement prediction, Finance Res. Lett. 58 (Part A) (2023) 104304.
[36] Yusheng Huang, Xiaoyan Mao, Yong Deng, Natural visibility encoding for time series and its application in stock trend prediction, Knowl.-Based Syst. 232 (2021) 107478.
[37] Shangzhe Li, Junran Wu, Xin Jiang, Ke Xu, Chart GCN: Learning chart information with a graph convolutional network for stock movement prediction, Knowl.-Based Syst. 248 (2022) 108842.
[38] Ruize Gao, Shaoze Cui, Hongshan Xiao, et al., Integrating the sentiments of multiple news providers for stock market index movement prediction: A deep learning approach based on evidential reasoning rule, Inform. Sci. 615 (2022) 529–556.
[39] Hongfeng Xu, Lei Chai, Zhiming Luo, Shaozi Li, Stock movement predictive network via incorporative attention mechanisms based on tweet and historical prices, Neurocomputing 418 (2020) 326–339.
[40] Hongfeng Xu, Donglin Cao, Shaozi Li, A self-regulated generative adversarial network for stock price movement prediction based on the historical price and tweets, Knowl.-Based Syst. 247 (2022) 108712.
[41] Jacob Devlin, Mingwei Chang, Kenton Lee, Kristina Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv:1810.04805.

[42] Yinhan Liu, Myle Ott, Naman Goyal, et al., Roberta: A robustly optimized bert pretraining approach, 2019, arXiv:1907.11692.
[43] Xingtong Chen, Xiang Ma, Hua Wang, Xuemei Li, Caiming Zhang, A hierarchical attention network for stock prediction based on attentive multi-view news learning, Neurocomputing 504 (2022) 1–15.
[44] Daniel Tang, Zhenghan Chen, Kisub Kim, Yewei Song, Haoye Tian, Saad Ezzini, Yongfeng Huang, Jacques Klein Tegawende F Bissyande, Collaborative agents for software engineering, 2024, arXiv preprint arXiv:2402.02172.
[45] M. Kachuee, H. Yuan, Y.B. Kim, S. Lee, Self-supervised contrastive learning for efficient user satisfaction prediction in conversational agents, 2020, arXiv preprint arXiv:2010.11230.
[46] W. Falcon, K. Cho, A framework for contrastive self-supervised learning and designing a new approach, 2020, arXiv preprint arXiv:2009.00104.
[47] Ilya Loshchilov, Frank Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, ICLR, 2018.
[48] Thomas Wolf, Lysandre Debut, Victor Sanh, et al., Huggingface' s transformers: State-of-the-art natural language processing, 2019, arXiv preprint: arXiv:1910.03771.
[49] E.F. Fama, K.R. French, International tests of a five-factor asset pricing model, J. Financ. Econ. 123 (3) (2017) 441–463.

Song Yang received her Ph.D. from Nanjing University, Nanjing, China, in 2020. She is now a teacher at the School of Mathematics and Statistics at Guizhou University of Finance and Economics. She is interested in forecasting financial time series, deep neural networks and corporate finance.



Daniel Tang is a doctoral researcher with a deep dive into various code-related areas. He delves into topics like patch representation, security patch detection, and pull request automation, to name a few. His expertise also spans object-level fault localization, bug repair, and object-level patch generation. Pushing the boundaries, Daniel has introduced frameworks for text generation that mimic unit code and has ventured into automated plugin creation. His dedication to reshaping the future of coding is evident. Professionally, he serves as a reviewer and pc member for esteemed conferences like ACL, NLP, SOSP, and more.