Astock: A New Dataset and Automated Stock Trading based on Stock-specific News Analyzing Model

Jinan Zou [*], **Haiyao Cao** [*], **Lingqiao Liu, Yuhao Lin, Ehsan Abbasnejad, Javen Qinfeng Shi** [†]

Australian Institute for Machine Learning, The University of Adelaide, Australia

`jinan.zou, haiyao.cao, lingqiao.liu@adelaide.edu.au`
`yuhao.lin01, ehsan.abbasnejad, javen.shi@adelaide.edu.au`

Abstract

Natural Language Processing(NLP) demonstrates a great potential to support financial decision-making by analyzing the text from social media or news outlets. In this work, we build a platform to study the NLP-aided stock auto-trading algorithms systematically. In contrast to the previous work, our platform is characterized by three features: (1) We provide financial news for each specific stock. (2) We provide various stock factors for each stock. (3) We evaluate performance from more financial-relevant metrics. Such a design allows us to develop and evaluate NLP-aided stock auto-trading algorithms in a more realistic setting. In addition to designing an evaluation platform and dataset collection, we also made a technical contribution by proposing a system to automatically learn a good feature representation from various input information. The key to our algorithm is a method called semantic role labeling Pooling (SRLP), which leverages Semantic Role Labeling (SRL) to create a compact representation of each news paragraph. Based on SRLP, we further incorporate other stock factors to make the stock movement prediction. In addition, we propose a self-supervised learning strategy based on SRLP to enhance the out-of-distribution generalization performance of our system. Through our experimental study, we show that the proposed method achieves better performance and outperforms all the baselines' annualized rate of return as well as the maximum drawdown of the CSI300 index and XIN9 index on real trading. Our Astock dataset and code are available at https://github.com/JinanZou/Astock.

1 Introduction

The stock prediction has been an attractive task for a long time, and it is still challenging since the stochasticity of the market and behavior patterns

Figure 1: Overview of the automated stock trading system.

of participators are fluctuating and elusive. Stock forecasting based on Natural Language Processing (NLP) techniques is a promising solution since text information, e.g., tweets, financial news etc., is strongly correlated with the stock prices. However, the NLP-based stock forecasting research is still scattered without unified definitions, benchmark datasets, clear articulations of the tasks, which severaly hinders progress of this field.

Existing approaches are usually based on market sentiment analysis (Xu and Cohen, 2018; Cheng and Li, 2021) and use news to predict the related securities' price on the following trading day(s) (Zhang et al., 2017; Li et al., 2020). Despite the limited success in those studies, the existing works are still far from realistic for two reasons: Firstly, previous methods ignore the financial factors, which plays a key role in practical trading. Secondly, these models are evaluated only on intermediate performance metric, e.g., stock movement prediction accuracy. It is unclear how well they can support a practical trading system to make sufficient profit.

To address the problems above, we construct a China A-shares market dataset with news and stock factors called Astock. Specifically, we an-

---

[*] Both authors contributed equally
[†] Corresponding Author

notate all occurrences of the three trading actions (long, preserve, short) in 40,963 news originated from Tushare [1] with a valid official license, which describes the major financial events. The dataset also includes various stock factors to build a realistic system. Based on Astock, we establish a semantic role labeling pooling (SRLP) to build a compact representation for stock-specific news and predict the stock movement. This work also explores how to leverage a self-supervised method better to upgrade the SRLP method, which achieves better performance for classification and high domain generalization ability.

In experiments, we further propose a realistic trading platform that outperforms the state-of-the-art text classification baseline's average returns and Sharpe Ratios over the CSI300 index and XIN9 index of testing period from January 2021 to November 2021. Specifically, we analyze the profitability of the proposed strategy based on stock movement prediction result for real trading as shown in Figure 1. The primary contributions of this work can be summarized as follows:

- We construct a brand new Chinese stock movement prediction task dataset with stock-specific news and stock factors.

- Our SRLP characterizes the key attributes of financial events, which is convenient for incorporating other stock factors and further creating a self-supervised module on top of the SRLP method. Our self-supervised SRLP method obtains competitive stock movement prediction and out-of-distribution (OOD) generalization results.

- We further evaluate algorithm performance on real-world trading from more financial-relevant metrics. By conducting extensive experimental studies, we show that our self-supervised SRLP achieves remarkable performance on these metrics. Furthermore, we observe that the proposed trading strategies work well in practice.

2 Related Work

2.1 Text-based Stock Movement Prediction

In recent years, the use of text-based information, especially news and social media, has significantly

improved the performance of stock movement prediction tasks and these methods usually rely on text-based features and sentiment analysis to forecast stock movements (Xu and Cohen, 2018; Hu et al., 2018; Ding et al., 2015). However, These approaches assume that the real-trading distribution was the same as the training distribution, which is not realistic as it is difficult to generalize to future trading. By contrast, our self-supervised SRL approach pays closer attention to the quality and comprehensiveness of the news, which could help with out-of-distribution generalization on the realistic trading.

2.2 Semantic Role Labeling and Self-Supervised Learning Approach

Semantic role labeling (SRL) aims to disclose the predicate-argument structure of a given sentence, which could provide a clear overlay that uncovers the underlying semantics of text (Conia et al., 2021). However, previous stock movement prediction methods (Xu and Cohen, 2018; Hu et al., 2018; Ding et al., 2015) adopted the word or sentence level representation to predict the stock movement. Due to the lack of abstract information of the news, these approaches can overfit the training data and fail to distinguish the key features of news. To deal with this problem, we used the SRL's characteristics for extracting a clear overlay that uncovers the underlying semantics of news.

Recently, self-supervised learning has become a very popular technique in the training stage of NLP, which generates labels without any human intervention and learns common language representations. Some researches (Im et al., 2021; Zheng et al., 2021) have proven that self-supervised learning strengthens the generalization ability for models as it improves the performance in many tasks.

3 Dateset Creation

Table 1: The comparison between Astock and other existing widely-used stock movement prediction dataset.

Dataset	Num of Stock	Text Source	Price-level	Stock Factors
DMFT's dataset (Zhang et al., 2017) EMNLP 17'	50	✗	Daily	✗
StockNet's dataset (Xu and Cohen, 2018) ACL 18'	88	Twitter	✗	✗
Dingxia's dataset (Ding et al., 2014) EMNLP 14'	500	News	Daily	✗
Trade the event 's dataset (Zhou et al., 2021) ACL 21'	✗	News	✗	✗
Ours	3680	Stock News	Minute-level when news published Daily-level for all the stocks	✔

The stock movement prediction task aims to explore a realistic method to predict the stock move-

ment with comprehensive and reasonable information in the China stock market. To this end, it is important to have minute-level price information in the dataset and we are motivated to collect one.

3.1 Standard of news and stock factors collection

There are two main components in our dataset: News and stock factors for the China stock market. In terms of news data, there are 40,963 pieces of listed company news, including company announcements and company-related news from July 2018 to November 2021. The news data are split into two parts: the In-distribution split and the out-of-distribution split. The in-distribution split is from July 2018 to December 2020 for training and testing where the training set occupies by 80%, and the validation set and test set occupy 10% respectively. The out-of-distribution split is selected from January 2021 to November 2021, which is used for OOD generalization testing. Every piece of news includes its published time and a corresponding news summary. Factor investing is an investment approach that involves targeting quantifiable firm characteristics or factors that can explain the differences in stock returns. Factor-based strategies may help investors meet particular investment objectives—such as potentially improving returns or reducing risk over the long term. Our Astock dataset covers the 24 stock factors on each stock of the China A-shares including Dividend yield, Total share, Circulated share, Free Float share, Market Capitalization, Price-earning ratio, PE for Trailing Twelve Months, Price/book value ratio, Price-to-sales Ratio, Price to Sales ratio, Circulate Market Capitalization, Open price, High price,Low price, Close price , Previous close price, Price change, Percentage of change, Volume, Amount, Turn over rate, Turn over rate for circulated Market Capitalization, Volume ratio. Furthermore, We compare Astock with several widely used stock movement prediction datasets in Table 1. The value is reflected in the following aspects: (1) Astock provides financial news for each specific stock over the entire China A-shares market. (2) Astock provides various stock factors for each stock. (3) Astock provides minute-level historical prices for the news.

3.2 Task Formulation

We divide the automated trading system into two tasks: stock movement classification and simulated trading.

3.2.1 Text-based stock movement classification

The goal of the stock movement classification task is to classify the effects of the input information. We measure the impact of each piece of company news by the stock return rate. In this paper, the news is annotated by the stock return rate r , and three cases are considered in our annotation: outperforming, neutral, and underperforming as shown in Equation 1. We further model the stock movement by classifying it into three categories. The ground truth for those categories can be derived from r. Specifically, we follow the following rules to categorize the data into three classes after ranking all the news by r, which aims to find the most strong signal of the stock movement, and to reduce the disturbance of noises comparing to dividing the data evenly. After the domain experts gave us the advice and the experiments with different thresholds was conducted, we set 20% as the threshold where the tunable parameters a, b, c, and d are 20, 40, 60, and 20, respectively.

$$\text{label=}\begin{cases} \text{outperforming,} & \text{if } r \text{ ranked top a\%} \\ \text{neutral,} & \text{if } r \text{ ranked top b\%-c\%} \\ \text{underperforming,} & \text{if } r \text{ ranked bottom d\%} \end{cases}$$
$$(1)$$

where r is the return rate of the news. We randomly select 80% of the in-distribution dataset as the training set, and the other 20% is split evenly into validation and test sets.

3.2.2 Simulated Trading

Stock movement prediction accuracy may not necessarily translate to a profitability of an auto-trading system. To further investigate how the stock movement prediction can benefit for the actual trading practice, we employ a practical trading strategy based on the stock movement prediction results and evaluate various metrics for the trading actions. The trading strategy details can be found at our github page.

4 Methodology

This section describes the technical contribution of this work: a novel system for stock movement prediction. Our system consists of two major components: semantic role labeling pooling method and a self-supervised learning based on SRLP, we will elaborate on those two parts.
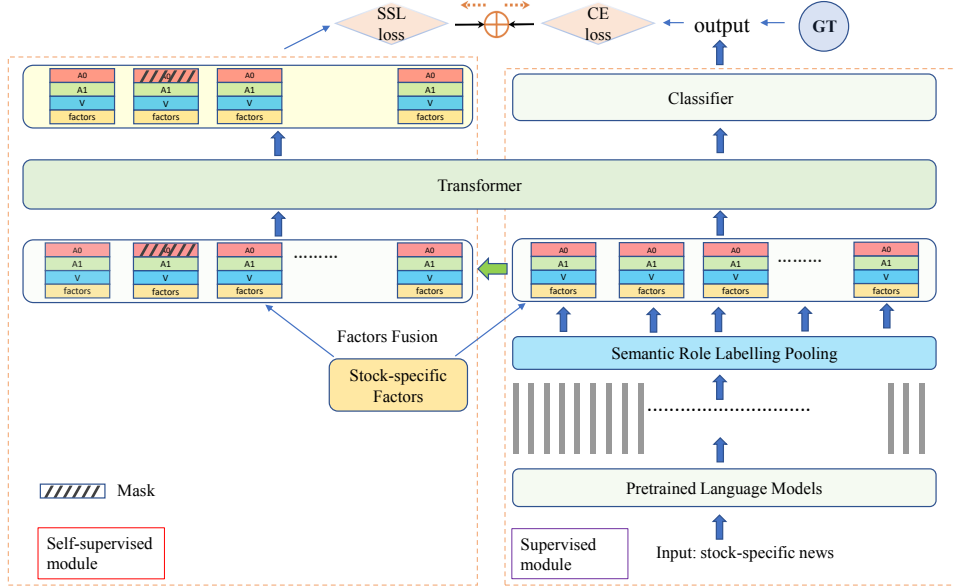
Figure 2: Overall framework of our approach, including a domain adapted pre-trained model (RoBERTa WWM Ext), Semantic Roles Pooling, transformer layer, self-supervised module (left part), and the supervised module (right part). The green arrow represents a duplicate for the SRLP. The final result is generated from the stock movement classifier, and the total loss is obtained from the self-supervised SRLP part and supervised stock movement classification part.

4.1 Semantic Role Labelling Pooling

In this work, we propose to leverage the off-the-shelf semantic role labeling, i.e., Propbank (Kingsbury and Palmer, 2003), to pool the output embeddings of a pre-trained language model to construct an alternative representation. The rationale is that the semantic roles in Propbank, i.e., verb (V), proto-agent (A0), and proto-patient (A1), are general-purposed and are also strongly associated with the event arguments. We show an example for semantic role labeling for financial news in Figure 3.

Original text: 科锐国际股东Career HK减持公司股份199万股,占公司总股本的1.103%。

Translation: Career HK, a shareholder of Kerui international, reduced 1.99 million shares of the company, accounting for 1.103% of the total share capital of the company

A0	V	A1
科锐国际股东Career HK	减持	公司股份199万股
Career HK, a shareholder of Kerui international	reduced	1.99 million shares

Figure 3: A Semantic role labeling example for a piece of news.

More specifically, we first use the Language Technology Platform (LTP) (Che et al., 2020) to automatically mark the semantic roles from the sentences of an entire piece of news and then select V, A0, and A1 to represent the roles for each sentence. Secondly, we process each sentence with a pretrained language model to obtain a sequence of output embeddings $\{\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_n\}$. We use $\mathcal{V}$, $\mathcal{A}_0$ and $\mathcal{A}_1$ to denote the indices of tokens corresponding to the V, A0, A1 components. At last, we perform pooling for embeddings with their indices falling into $\mathcal{V}$, $\mathcal{A}_0$ and $\mathcal{A}_1$. We call this scheme Semantic Role Labelling Pooling SRLP in short. Taking A0 as an example, the SRLP feature for A0 is

$$\mathbf{e}_{A0} = \frac{1}{|\mathcal{A}_0|} \sum_{i \in \mathcal{A}_0} \mathbf{s}_i \qquad (2)$$

For a sentence with N sets of V, A0 and A1, we concatenate $\mathbf{e}_{A0}, \mathbf{e}_{A1}, \mathbf{e}_V$ of each sentence and the financial factor $\mathbf{F}$ of the stock-of-interest into a data matrix:

$$\mathbf{E} = \begin{bmatrix} \mathbf{e}_V^1 & \dots & \mathbf{e}_V^t & \dots & \mathbf{e}_V^N \\ \mathbf{e}_{A0}^1 & \dots & \mathbf{e}_{A0}^t & \dots & \mathbf{e}_{A0}^N \\ \mathbf{e}_{A1}^1 & \dots & \mathbf{e}_{A1}^t & \dots & \mathbf{e}_{A1}^N \\ \mathbf{F} & \dots & \mathbf{F} & \dots & \mathbf{F}, \end{bmatrix} \qquad (3)$$

where $\mathbf{E}$ is output of the above process. Each column of $\mathbf{E}$, denoted as $\mathbf{e}_j$, is the concatenation of $\mathbf{e}_V^j, \mathbf{e}_{A0}^j, \mathbf{e}_{A1}^j$ and $\mathbf{F}$. $\mathbf{E}$ is then processed by a Transformer encoder in the same way as the standard text classification to generate the stock movement prediction.

4.2 Self-Supervised Learning based on SRLP

Besides standard supervised training loss for stock movement classification, in this work, we further propose to use a self-supervised training task as an auxiliary task to train the network. For stock movement prediction, good generalization is highly desirable since the training data is usually sampled from a period different from the test period. A significant problem in practice is to ensure that our model generalizes to scenarios different from the training set. We further create a self-supervised learning method on top of the SRLP. Recent studies (Mohseni et al., 2020; Hendrycks et al., 2019) have shown that incorporating a self-supervised learning task along with the supervised training task could lead to better generalization. As shown in Figure 2, the self-supervised task is defined as predicting the position of one randomly masked SRL role from all the roles of SRL in a piece of news. Intuitively, the self-supervised learning task should be designed to encourage the favorable properties of features. In this work, we propose to randomly mask one pooled embedding, i.e., $\mathbf{e}_V^j$, $\mathbf{e}_{A0}^j$ or $\mathbf{e}_{A1}^j$, from a randomly selected sentence, and then ask the network to identify the masked embedding from a pool of candidate embeddings. Such a cloze-style task encourages the network to perform reasoning over other unmasked cues to work out the missing item. We hypothesize that such a reasoning capability is beneficial for understanding the financial news and thus helps stock movement prediction.

Formally, we randomly select a $\mathbf{e}_j$ from $\mathbf{E}$ and then select one element from $\mathbf{e}_j = \{ \mathbf{e}_V^j, \mathbf{e}_{A0}^j, \mathbf{e}_{A1}^j \}$, after that we replace the selected element with an all-zero vector, indicating a "mask" operation. Taking masked V at the t-th sentence as an example, we denote the $\mathbf{E}$ after this mask operation as $\mathbf{E}'$.

$$\mathbf{E}' = \begin{bmatrix} \mathbf{e}_V^1 & ... & \mathbf{M} & ... & \mathbf{e}_V^N \\ \mathbf{e}_{A0}^1 & ... & \mathbf{e}_{A0}^t & ... & \mathbf{e}_{A0}^N \\ \mathbf{e}_{A1}^1 & ... & \mathbf{e}_{A1}^t & ... & \mathbf{e}_{A1}^N \\ \mathbf{F} & ... & \mathbf{F} & ... & \mathbf{F} \end{bmatrix}$$

Then we feed $\mathbf{E}'$ into the transformer to obtain a query vector sequence $\mathbf{q} \in \mathbb{R}^d$

$$\mathbf{q} = Transformer(\mathbf{E}')[:, t]$$

where $[:, t]$ means extract the t-th column of the vector sequences calculated by the transformer. The unmasked SRLP-V features (or SRLP-A0, SRLP-A1 features, depending on which type of SRLP

feature is chosen) is also send to an encoder to calculate candidate key vectors: Formally, $\mathbf{K}$ is defined as:

$$\mathbf{K} = [f_V(\mathbf{e}_V^1), \cdots, f_v(\mathbf{e}_V^t), \cdots, f_V(\mathbf{e}_V^N)] \in \mathbf{R}^{d \times N}$$

where f_V is an encoder specified for encoding V-type SRLP feature. Then the query vector is compared against each column vector in $\mathbf{K}$ and is expected to have the highest matching score at the t-th location. This process could be implemented via matrix multiplication and the softmax operation:

$$P_{SSL} = \text{Softmax}(\mathbf{qK}) \qquad (4)$$

and we hope the highest probability entry in Eq. 4 is at the t-th dimension. This requirement could be enforced via the cross-entropy loss. Finally, the training loss for the models is

$$\mathcal{L} = \alpha \mathcal{L}_{CLS} + (1 - \alpha) \mathcal{L}_{SSL} \qquad (5)$$

where $\mathcal{L}_{CLS}$ is the cross-entropy loss for the text classification and $\mathcal{L}_{SSL}$ is the cross-entropy loss on the self-supervised learning prediction P_{SSL}. α here is a trade-off parameter.

5 Experiments

In this section, we conduct experiments to evaluate the performance of the proposed model. We conduct experiments on two different splits of our dataset for each model: In-distribution split and out-of-distribution split. We also feed the prediction result of our method to the proposed trading strategy to analyze the profitability through back-testing on real-world stock data.

5.1 Evaluation metrics

We evaluated the stock movement prediction and simulated trading performance. For the Stock movement prediction, we applied the **Accuracy, F1 Score, Recall and Precision** as evaluation metrics. For simulated trading, we applied the **Annualized Rate of Return, Maximum Drawdown and Sharpe Ration** as evaluation metrics based on our simulated trading strategy.

5.2 Compared Methods

We re-implement the current state-of-art stock movement prediction models as baselines, including StockNet (Xu and Cohen, 2018), HAN Stock (Hu et al., 2018). **StockNet (Xu and Cohen, 2018)**

Table 2: The stock movement classification performance(%) of in-distribution evaluation on our scheme and others demonstrates the effectiveness of our self-supervised SRL method. ✔ indicates that the model adopted this Semantic role's pooling information. **-** indicates that the method does not adopt this semantic role's pooling. ✗ indicates that semantic role's pooling is masked.

Model	Resource	Semantic Role			Accuracy	F1 Score	Recall	Precision
		A0	V	A1				
StockNet (Xu and Cohen, 2018) ACL 18'	News	-	-	-	46.72	44.44	46.68	47.65
HAN Stock (Hu et al., 2018) ICWSDM 18'	News	-	-	-	57.35	56.61	57.20	58.41
Bert Chinese (Devlin et al., 2019)NAACL 19'	News	-	-	-	59.11	58.99	59.20	59.07
ERNIE-SKEP (Tian et al., 2020) ACL 20'	News	-	-	-	60.66	60.66	60.59	61.85
XLNET Chinese (Cui et al., 2020)EMNLP 20'	News	-	-	-	61.14	61.19	61.09	61.60
RoBERTa WWM Ext (Cui et al., 2020)EMNLP 20'	News	-	-	-	61.34	61.48	61.32	61.97
	News + Factors	-	-	-	62.49	62.54	62.51	62.59
Our SRLP	News	✔	✔	✔	61.76	61.69	61.62	61.87
	News + Factors	✔	✔	✔	64.79	64.85	64.79	65.26
Our Self-supervised SRLP	News	✗	✔	✗	61.07	61.11	61.11	61.11
	News	✗	✔	✔	62.36	62.32	62.43	62.64
	News	✔	✔	✗	62.42	62.46	62.44	62.62
	News	✗	✗	✔	62.15	62.15	62.15	62.59
	News	✔	✗	✗	61.34	61.23	61.46	61.30
	News	✔	✗	✔	62.97	63.05	62.93	63.47
Our Self-supervised SRLP with Factors	News + Factors	✗	✔	✗	64.59	64.62	64.63	64.65
	News + Factors	✗	✔	✔	66.82	66.81	66.90	66.82
	News + Factors	✔	✔	✗	65.54	65.53	65.62	65.50
	News + Factors	✗	✗	✔	65.34	65.21	65.43	65.43
	News + Factors	✔	✗	✗	65.27	65.35	65.24	65.77
	News + Factors	✔	✗	✔	**66.89**	**66.92**	**66.95**	**66.92**

is a stock temporally-dependent movement prediction model which also uses Twitter data and price information to predict two classes for stock movement. Hybrid Attention Networks **(HAN) Stock** (Hu et al., 2018) is a stock trend prediction model based on a sequence of recent related news to predict three classes for stock movement task, which are the same as ours. We also construct baselines by formulating the stock movement prediction problem as text classification and use four strong pre-trained Chinese language models as backbones such as **XLNet-base-Chinese**(Cui et al., 2020), Sentiment Knowledge Enhanced pre-trained language model(**SKEP**)(Tian et al., 2020), **Bert Chinese**(Devlin et al., 2019) and **RoBERTa WWM Ext**(Cui et al., 2020). For the above four pre-trained language models, we extract sentence embedding from the [CLS] token and attach a three-way classifier to predict the stock movements. In addition, we also compare the CSI300 index, XIN9 index[2] against the proposed method when analyzing the profitability of the proposed system. For our methods, we used RoBERTa WWM Ext as our

backbone PLM since the remarkable performance.

5.3 Stock Movement Evaluation

We first compare different methods on the task of stock movement prediction. We conduct experiments on two different splits of our dataset: In-distribution split and out-of-distribution split. In the in-distribution split, both training and testing data are sampled from the same period while the out-of-distribution split uses data from different periods to construct the training and testing data.

In-distribution evaluation

The results are shown in Table 2. From the results, we made the following two observations:

1. If only text information is used, the proposed SRLP approach achieves the state-of-the-art performance. Interestingly, we find that SRLP achieves superior performance when further combines the stock factors. It outperforms RoBERTa WWM Ext (News+Factors) by more than 2%. We postulate that this is because the compact representation in SRLP make incorporation of stock factors easier. Note that the proposed way of incorporating stock factors (see Section 3.1) does not only introduce extra modalities for the stock movement prediction but also could make the text analysis module

---

[2]Equivalent to the Standard and Poor's 500 (S&P 500) or the Dow Jones Industrial Average (DJIA) in the US stock market

adaptive to the stock factors. This could be useful to model the scenario like the effect of a similar event could result in a different impact on the stock movement for a different type of company.

2. The proposed self-supervised SRLP can further boost the performance of SRLP. In the best setting of self-supervised SRLP, i.e., with V being masked, self-supervised SRLP achieves more than 1% improvement over SRLP. The improvement is even larger when the stock factors are provided, showing more than 2% improvement over SRLP (News+Factors), achieving 66.89% prediction accuracy. This validates the effectiveness of the proposed self-supervised learning approach. Interestingly, we observe that masking A0 and A1 usually will not bring improvement in contrast to the case of masking V. Note that V encodes the type of an event, and the argument is encoded by A0 or A1. It seems that predicting the type of events is a more effective self-supervised learning task than working on the argument.

Out-of-distribution evaluation

In the experiments above, the training data and testing data are sampled from the same period. Thus the distributions of training data and testing data are similar. For real-world applications, the stock movement prediction model is applied to future data unseen at the training time. Hence, it is critical to evaluate the model in such an out-of-distribution setting.

Table 3: The comparison (%) of the out-of-distribution evaluation on stock movement classification with Stock-Net, RoBERTa-WWM Ext, HAN Stock method and our method from 1/1/2021 to 12/11/2021.

Model	Accuracy	F1 Score	Recall	Precision
StockNet(Xu and Cohen, 2018)	44.35	42.52	45.42	45.82
HAN Stock(Hu et al., 2018)	53.41	53.33	53.69	54.53
RoBERTa WWM Ext(Cui et al., 2020)	60.15	60.08	59.89	60.78
Ours	**64.09**	**63.95**	**63.90**	**64.43**

To this end, we construct a new training/testing split by using the data from July 2018 to December 2020 as training data and the data from January 2021 to November 2021 for the testing data. We first conduct an evaluation on the stock movement prediction task, and the results are shown in Table 3. From the results, we can see that the proposed method is still comparably competitive over other baselines.
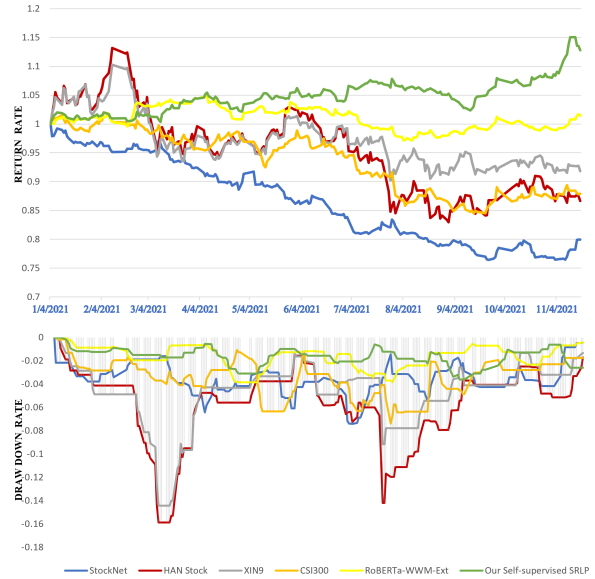

Figure 4: The comparison for the real trading performance on Return Rate, Draw Down Rate with CSI300 index, XIN9, Roberta WWM Ext, HAN Stock, StockNet and our proposed method from 1/1/2021 to 12/11/2021

5.4 Profitability Evaluation in Real-world

Table 4: The comparison of profitability test on Maximum Drawdown(%), Annualized Rate of Return(%), and Sharpe Ratio Rate(%) with strong baselines, XIN9, CSI300 and our proposed method from 1/1/2021 to 12/11/2021.

Model	Maximum Drawdown↓	Annualized Rate of Return↑	Sharpe Ratio↑
XIN9	-15.85	-15.38	-32.01
CSI300	-14.40	-9.34	-32.99
StockNet(Xu and Cohen, 2018)	-7.40	-22.42	-177.65
HAN Stock(Hu et al., 2018)	-7.38	-13.50	-55.84
RoBERTa WWM Ext(Cui et al., 2020)	-3.83	1.35	-16.31
Ours	**-3.60**	**13.85**	**40.93**

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 January 2021 to November 2021 using the stock movement prediction result of our model trained from July 2018 to December 2020 as mentioned in Section 3.1. In Table 4, we show that our self-supervised SRLP model achieves a remarkable annualized rate of return of 13.85% , which surpasses the previous baselines and market index XIN9 and CSI300. The resulting baseline HAN Stock (Hu et al., 2018) and StockNet (Xu and Cohen, 2018) achieve an annualized rate of return of -13.5% and -22.42% re-

spectively, and the market XIN9 index and CSI300 were overall declining in 2021, which obtains -15.38 % and -9.34% respectively. In addition, our self-supervised learning method also obtains the lowest Maximum Drawdown of -3.6% and the highest Sharpe Ratio of 40.93% , which significantly outperforms the previous methods and indicates that our self-supervised could successfully achieve higher expected returns while remaining relatively less risky as shown in Figure 4.

6   Conclusion

In this paper, we study the problem of NLP-based stock movement prediction and build a platform with a new dataset, AStock, featured by: (1) Large number of stocks, stock-relevant news. (2) Availability of various financial factors. (3) Financial-relevent metrics for Evaluation. The Platform is based on two novel techniques. One leverages Propbank-style semantic role labeling results to create compact news representation. Building on top of this representation, the other technique is a customized self-supervised learning training strategy for improving generalization performance. We demonstrate that the proposed method achieves superior performance over other baselines through extensive experiments in both in-distribution and out-of-distribution settings. Also, by feeding our prediction to a practical simulated trading, our method achieves better profitability in backtesting.

References

Wanxiang Che, Yunlong Feng, Libo Qin, and Ting Liu. 2020. N-ltp: A open-source neural chinese language technology platform with pretrained models. *arXiv preprint arXiv:2009.11616*.

Rui Cheng and Qing Li. 2021. Modeling the momentum spillover effect for stock prediction via attribute-driven graph attention networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 55–62.

Simone Conia, Riccardo Orlando, Fabrizio Brignone, Francesco Cecconi, and Roberto Navigli. 2021. Invero-xl: Making cross-lingual semantic role labeling accessible with intelligible verbs and roles. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), Punta Cana, Dominican Republic*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

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

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2014. Using structured events to predict stock price movement: An empirical investigation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1415–1425, Doha, Qatar. Association for Computational Linguistics.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep learning for event-driven stock prediction. In *Twenty-fourth international joint conference on artificial intelligence*.

Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. 2019. Using self-supervised learning can improve model robustness and uncertainty. *Advances in Neural Information Processing Systems (NeurIPS)*.

Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. 2018. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 261–269.

Jinbae Im, Moonki Kim, Hoyeop Lee, Hyunsouk Cho, and Sehee Chung. 2021. Self-supervised multimodal opinion summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 388–403, Online. Association for Computational Linguistics.

Paul Kingsbury and Martha Palmer. 2003. Propbank: the next level of treebank. In *Proceedings of Treebanks and lexical Theories*, volume 3. Citeseer.

Yelin Li, Hui Bu, Jiahong Li, and Junjie Wu. 2020. The role of text-extracted investor sentiment in chinese stock price prediction with the enhancement of deep learning. *International Journal of Forecasting*, 36(4):1541–1562.

Sina Mohseni, Mandar Pitale, JBS Yadawa, and Zhangyang Wang. 2020. Self-supervised learning for generalizable out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5216–5223.

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and feng wu. 2020. SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4067–4076, Online. Association for Computational Linguistics.

Yumo Xu and Shay B. Cohen. 2018. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1979, Melbourne, Australia. Association for Computational Linguistics.

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

Yuanhang Zheng, Zhixing Tan, Meng Zhang, Mieradilijiang Maimaiti, Huanbo Luan, Maosong Sun, Qun Liu, and Yang Liu. 2021. Self-supervised quality estimation for machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3322–3334.

Zhihan Zhou, Liqian Ma, and Han Liu. 2021. Trade the event: Corporate events detection for news-based event-driven trading. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2114–2124, Online. Association for Computational Linguistics.

A Appendix

A.1 Stock Financial Factors

The stock financial factors in our dataset include Fundamental factors: Dividend yield, Total share, Circulated share, Free circulated share, Market Capitalization, Price-earning ratio(PE), PE for Trailing Twelve Months(TTM), Price/book value ratio, Price-to-sales Ratio, Price to Sales ratio (TTM), Circulate Market Capitalization, Open price, High price,Low price, Close price , Previous close price, Price change, Percentage of change, Volume, Amount, Turn over rate, Turn over rate for circulated Market Capitalization, Volume ratio.