Yet at the FinNLP-2022 ERAI Task: Modified models for evaluating the Rationales of Amateur Investors

Yan Zhuang
University Of Electronic Science And Technology Of China
delecisz@gmail.com

Fuji Ren*
University Of Electronic Science And Technology Of China
renfuji@uestc.edu.cn

Abstract

The financial reports usually reveal the recent development of the company and often cause the volatility in the company's share price. The opinions causing higher maximal potential profit and lower maximal loss can help the amateur investors choose rational strategies. FinNLP-2022 ERAI task aims to quantify the opinions' potentials of leading higher maximal potential profit and lower maximal loss. In this paper, different strategies were applied to solve the ERAI tasks. Valinna 'RoBERTa-wwm' showed excellent performance and helped us rank second in 'MPP' label prediction task. After integrating some tricks, the modified 'RoBERTa-wwm' outperformed all other models in 'ML' ranking task.

1 INTRODUCTION

With the development of data mining and natural language processing techniques, more and more people are looking at textual information in various fields. One such area is finance. Based on the financial corpora, researchers have pre-trained several models, like Mengzi-Fin (Zhang et al., 2021) and various versions of FinBERT (Liu et al., 2021; Yang et al., 2020; Araci, 2019), which help better learn the semantic layer of financial domain knowledge and more comprehensively learn the feature distribution of financial domain words and phrases. Besides, there are a number of researchers who predict future events with texts (Zong et al., 2020), like mining the sentiment of financial posts to predict which stock has better returns (Chen et al., 2021b). Chen et al. (2021a) compares the rationales of experts and those of the crowd from stylistic and semantic perspectives to find the top-ranked opinions, and find they can increase potential returns and reduce downside risk.

In addition, FinNLP teams holds a series of workshops to help collect the research related to AI in FinTech (Chen et al., 2019, 2018; Zong et al., 2020; Chen et al., 2020), and to handle some frontier financial problems. This year they have partnered with EMNLP and hold ERAI shared task to evaluate the rationales of amateur investors by predicting the maximal potential profit (MPP) and maximal loss (ML) of the given analytical opinions (Chen et al., 2022). We participated it and came up with several solutions like changing the optimizer, using 'Stochastic Weight Averaging' method, which helped us rank 2nd in the 'MPP' classification subtask and 1st in the 'ML' ranking subtask.

2 TASK SETTING AND DATASETS

There are two subtasks in the Evaluating the Rationales of Amateur Investors (ERAI) shared task (Chen et al., 2022), namely 'Pairwise Comparison' and 'Unsupervised Ranking'. The former one includes two binary classification tasks. One aims to determine, given the opinion pairs, whether the given opinion 1 will lead to higher maximal potential profit (MPP) than the given opinion 2, while another requires to determine whether the opinion 1 will to higher maximal loss (ML) than the given opinion 2. 'Unsupervised Ranking' task requires to find out the top 10% of the given posts that will lead to higher MPP. The datasets are collected from one of the largest financial social media platforms in Taiwan, PTT Stock [1] and MObile01 (Chen et al., 2022). And the posts are available in both English and Chinese. There are 200 post pairs and their corresponding 'MPP' values and 'ML' values in the training phase, while 87 post pairs in testing phase of 'Pairwise Comparison' task and 210 posts in testing phase of 'Unsupervised Ranking' task.

*Corresponding author

[1] https://www.ptt.cc/bbs/Stock/index.html

Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP), pages 132 - 135

3 METHOD

We applied different strategies to handle the ERAI tasks. In both subtasks, we used a BERT-type pre-trained model (Devlin et al., 2018), but we treated 'Pairwise Comparison' subtask as a sentence pair classification task and 'Unsupervised Ranking' subtask as a regression task.

We processed Chinese posts in one more step than English posts, i.e. turning Traditional Chinese into Simplified Chinese using 'zhconv' library [2] for that many models were pre-trained on simplified Chinese corpus (Cui et al., 2021). Then we removed the '\n' characters, urls, and Emoticons in the posts. Finally, we used a max length of 128 truncation of the posts and fed the cleaned posts into the models.

3.1 Models for Pairwise Comparison Subtask

We used only the original pre-trained model in this task, in both Chinese and English, and applied three-fold cross validation for model fusion. Models include but are not limited to:

FinBERT [3] incorporated knowledge from the financial domain, introduced phrase and semantic level tasks, extracted proper nouns or phrases from the domain, and was pre-trained using a full word mask and two types of supervised tasks on Chinses corpora in the BERT pre-training procedure (Devlin et al., 2018).

Mengzi-Fin (Zhang et al., 2021) was pre-trained on financial news, announcements, research reports crawled from the web following RoBERTa pre-training procedure (Liu et al., 2019).

RoBERTa-large-pair (Xu et al., 2020) was prt-trained on CLUECorpus2020 using a semantic similarity model. It has a high probability of working better than using a direct pre-trained model in semantic similarity or sentence pair problems.

RoBERTa-wwm (Cui et al., 2021) was pre-trained on Chinese corpora using whole word masking (WWM). Points to note that the model is not the original RoBERTa model, but only a BERT model trained in a similar way to RoBERTa training, i.e. RoBERTa-like BERT.

3.2 Models for Unsupervised Ranking Subtask

We used the values of the 'MPP' and 'ML' columns corresponding to the two posts in 'Pairwise Comparison' subtask as targets to train the regression model. Three strategies were applied.

BERT-LR fed the features of [CLS] token from the BERT-base model into a regression layer, which consists of a dropout layer and linear layer. The model updated the weights of BERT-model and the regression layer.

BERT-lightGBM selected BERT-base model as the feature extractor, and put the selected features from the [CLS] token into lightGBM regressor. It is important to note that the model only updated the weights of lightGBM and not the weights of BERT.

Modified-RoBERTa-wwm chose RoBERTa-wwm as the backbone and modified it with 'Stochastic Weight Averaging' (SWA) (Izmailov et al., 2018), 'MADGRAD Optimizer' (Defazio and Jelassi, 2022) and multi-sample dropout (Inoue, 2019). Specifically, SWA generates an aggregate by combining the weights of the same network at different training stages, and then uses this model with the combined weights to make predictions. Here we trained the first 7 out of 10 epochs with learning rate 2e-5, and trained the left 3 epochs with learning rate 1e-4. Besides, we replaced the Adaw optimizer with MADGRAD optimizer for that the latter one showed excellent performance on deep learning optimization. Then the [CLS] token of all hidden states were averaged for multi-sample dropout, and the output were averaged for the final predicting.

4 EXPERIMENTS AND RESULTS

Seven models were adapted in 'Pairwise Comparison' task and accuracy was selected as the evaluation metric, while three strategies were applied in 'Unsupervised Ranking' task and average MPP and ML are used as the evaluation metric, just as table 1 and table 2 show. The definition and the calculation method of MPP and ML can be found in Chen et al. (2021a).

4.1 Experiments and Results on Pairwise Comparison Subtask

To better compare the effectiveness of each model, we first split the data into three folds and then trained the three models accordingly. The offline evaluation metric was the average accuracy of the three models. All the seven models we used in 'Pairwise Comparison' task shared a fixed training config. They were all trained for 3 epochs with

[2] https://pypi.org/project/zhconv/

[3] https://github.com/valuesimplex/FinBERT

Models	MPP Offline	MPP Online	ML Offline	ML Online
FinBERT	55.48	-	**59.01**	40.23
Mengzi-Fin	61.48	-	**59.01**	40.23
BERT-en	62.98	57.47	58.05	40.23
RoBERTa-en	60.48	-	58.47	-
RoBERTa-large-pair	63.00	57.47	-	-
RoBERTa-wwm	**63.98**	57.47	58.47	-
RoBERTa-large	57.53	-	-	-

Table 1: The evaluation metric is accuracy. '-' denotes that we don't test the corresponding model. The figures in 'MPP Offline' and 'MPP Online' columns are the averaged validation accuracy and test accuracy of the three-fold models in 'MPP' label prediction task repectively, and the highest accuracy is highlighted in boldface.

Models	Average MPP of Top 10% Posts	Average ML of Top 10% Posts
BERT-LR	8.52%	-4.35%
BERT-lightGBM	12.10%	-5.77%
Modified-RoBERTa-wwm	**14.61%**	**-3.24%**
Baseline	17.61%	-2.46%

Table 2: The evaluation metric was the average MPP and ML of the top 10% posts. Values in 'Average ML of Top 10% Posts' column are all negative may because the given golden label values are all negative. The best performance is highlighted in boldface and the baseline scores are underlined.

learning rate 4e-5, max input length 128, weight decay rate 0.01 and the Adam parameter 1e-8. Table 1 shows the offline and online performance different models. 'FinBERT', 'Mengzi-Fin', 'RoBERTa-large-pair' and 'RoBERTa-wwm' were trained on the Chinese posts, while the others were all trained on the English opinions. Although 'FinBERT' and 'Mengzi-Fin' were pre-trained on financial domain texts, they still performed worse than the models pre-trained on general domain corpora like 'BERT-en'. And the models pre-trained on Chinese corpora showed better performance than the ones pre-trained on English corpora. This may be because the English posts were translated and the translation can lead to errors, in addition to the fact that there are inherent differences between different languages. 'RoBERTa-wwm' achieved the best accuracy, which ranked 2nd in the MPP prediction task. However, all three models we submitted showed same accuracy on the test set, which may imply we should not split the dataset into three folds, or there is gap between the training and test dataset and our model don't learn anything.

4.2 Experiments and Results on Unsupervised Ranking Subtask

The training config of the models in 'Unsupervised Ranking' task were not the same. 'BERT-LR' was trained for 5 epochs with learning rate 4e-5 and max input length 300 while 'Modified-RoBERTa-wwm' was trained for 10 epochs with max input length 256. Besides, the first 7 epochs were trained with learning rate 2e-5 and the last 3 epochs with learning rate 1e-4. The [CLS] token of all hidden states were averaged and then put softmax layer, normalization layer, regressor with multi-sample dropout sequentially. Finally, the average output were used to make predictions. The baseline only used stylistic and semantic features of the posts, which can be found in Chen et al. (2021a).

The performance of the three models could be seen in table 2. The performance of all our models don't exceed the baseline. 'Modified-RoBERTa-wwm' outperformed the left two models in both tasks, while 'BERT-LR' performed worst in 'MPP ' rank subtask and second worse in 'ML' subtask. It is important to notice that 'Modified-RoBERTa-wwm' ranked first in all competition teams in 'ML' rank subtask. Due to time constraints, we did not apply either the ablation study or the model from the 'unsupervised ranking' task to the 'pairwise comparison' task, which may also be a good solution.

5 CONCLUSION

In this work, we introduced our system models in FinNLP-2022 ERAI task. In 'Pairwise Comparison' task, seven models were discussed

and 'RoBERTa-wwm' outperformed other models and helped us rank 2nd in the 'MPP' classification among all submissions. While in 'Unsupervised Ranking' task, we tried three strategies and 'Modified-RoBERTa-wwm', which incorporated 'Stochastic Weight Averaging' (SWA), 'MADGRAD Optimizer' and multi-sample dropout, showed best performance and ranked 1st in the 'ML' ranking subtask.

In the future, we want to apply the models in 'Unsupervised Ranking' task to 'Pairwise Comparison' task through predicting the 'MPP' and 'ML' values of the posts. Besides, we found that the values of 'MPP' and 'ML' showed a negative correlation in both 'Pairwise Comparison' task and 'Unsupervised Ranking' task. This may be because the 'MPP' values are all positive and the 'ML' values are all negative, and we are trying to figure out if this is the reason or not.

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2018. Ntusd-fin: a market sentiment dictionary for financial social media data applications. In *Proceedings of the 1st Financial Narrative Processing Workshop (FNP 2018)*, pages 37–43.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. Issues and perspectives from 10,000 annotated financial social media data. In *Proceedings of The 12th language resources and evaluation conference*, pages 6106–6110.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021a. Evaluating the rationales of amateur investors. In *Proceedings of the Web Conference 2021*, WWW '21, page 3987–3998, New York, NY, USA. Association for Computing Machinery.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021b. *From opinion mining to financial argument mining*. Springer Nature.

Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2019. Proceedings of the first workshop on financial technology and natural language processing. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*.

Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2022. Overview of the finnlp-2022 erai task: Evaluating the rationales of amateur investors. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Aaron Defazio and Samy Jelassi. 2022. Adaptivity without compromise: a momentumized, adaptive, dual averaged gradient method for stochastic optimization. *Journal of Machine Learning Research*, 23:1–34.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hiroshi Inoue. 2019. Multi-sample dropout for accelerated training and better generalization. *arXiv preprint arXiv:1905.09788*.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2018. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2021. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4513–4519.

Liang Xu, Xuanwei Zhang, and Qianqian Dong. 2020. Cluecorpus2020: A large-scale chinese corpus for pre-training language model. *arXiv preprint arXiv:2003.01355*.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.

Zhuosheng Zhang, Hanqing Zhang, Keming Chen, Yuhang Guo, Jingyun Hua, Yulong Wang, and Ming Zhou. 2021. Mengzi: Towards lightweight yet ingenious pre-trained models for chinese. *arXiv preprint arXiv:2110.06696*.

Shi Zong, Alan Ritter, and Eduard Hovy. 2020. Measuring forecasting skill from text. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5317–5331, Online. Association for Computational Linguistics.