Overview of the FinNLP-2022 ERAI Task: Evaluating the Rationales of Amateur Investors

Chung-Chi Chen,[1] Hen-Hsen Huang,[2] Hiroya Takamura,[1] Hsin-Hsi Chen[3]
[1] AIST, Japan
[2] Institute of Information Science, Academia Sinica, Taiwan
[3] Department of Computer Science and Information Engineering
National Taiwan University, Taiwan
c.c.chen@acm.org, hhhuang@iis.sinica.edu.tw,
takamura.hiroya@aist.go.jp, hhchen@ntu.edu.tw

Abstract

This paper provides an overview of the shared task, Evaluating the Rationales of Amateur Investors (ERAI), in FinNLP-2022 at EMNLP-2022. This shared task aims to sort out investment opinions that would lead to higher profit from social platforms. We obtained 19 registered teams; 9 teams submitted their results for final evaluation, and 8 teams submitted papers to share their methods. The discussed directions are various: prompting, fine-tuning, translation system comparison, and tailor-made neural network architectures. We provide details of the task settings, data statistics, participants' results, and fine-grained analysis.

1 Introduction

In the financial market, people have different reasons to make trading/investment decisions. Thanks to the development of social media platforms, people can share these reasons and discuss them with others rapidly. However, there are hundreds of thousands of posts on social media platforms every day. Selecting the posts (opinions) that have the potential to help investors make profitable investment decisions becomes a challenge. Inspired by the ideas of persuasive essay scoring (Ghosh et al., 2016) and argument quality assessment (Skitalinskaya et al., 2021; Hasan et al., 2021), we proposed a new task: evaluating investment opinions based on the rationales in the post (Chen et al., 2021).

There are some steps when reading and evaluating investment opinions. First, as in most sentiment analysis studies (Chen et al., 2020; Xing et al., 2020), investors need to identify the sentiment of the opinion (bullish/bearish/neutral). Second, investors will read the reasons that are provided to support the sentiment. Third, investors will evaluate whether these reasons are rational, and further decide whether to follow the suggestions in the opinion. When we attempt to select useful investment opinions automatically, we think that systems also need to follow the above steps. However, in many cases, it is hard to decide the ground truth for the opinion quality because it is somehow subjective and varies due to the viewpoints. In the debate scenario, we can use the voting records as a proxy for evaluation. In the financial market, we can use historical information as a proxy to assess forecasting skills (Zong et al., 2020). Therefore, we propose to use maximum possible profit (MPP) and maximum loss (ML) as evaluation metrics to measure the quality of investment opinions (Chen et al., 2021).

In this shared task, we propose two kinds of settings, pairwise comparison and unsupervised ranking. The findings under these settings not only can be used in investment recommendations in the future, but also can be used in evaluating the generated reports and investor education. Additionally, we also expect that we can improve models' performances in market information forecasting tasks by sorting out high-quality opinions and filtering out low-quality opinions in the first step when selecting input data. Participants explore various directions for solving these challenges. There are several interesting discussions for a better understanding of where we are in the financial opinion scoring. We summarize the details of their methods in Section 3.

2 Tasks and Datasets

2.1 Task Setting

In ERAI shared task, we use MPP and ML to label opinions. Below are the definitions of MPP and ML in our previous work (Chen et al., 2021):

$$MPP_{bullish} = (\max(H_{(t+1,T)}) - O_{t+1})/O_{t+1} \quad (1)$$

$$ML_{bullish} = (\min(L_{(t+1,T)}) - O_{t+1})/O_{t+1} \quad (2)$$

Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP), pages 99 - 103

Team	Language Model	Method & Features & Lexicon
PromptShots (Wiriyathammabhum, 2022)	T5-Small (Raffel et al., 2020) Instruct-GPT (Ouyang et al., 2022) text-davinci-002 FinBERT-tone (Yang et al., 2020)	Part of Speech FinProLex (Chen et al., 2021) NTUSD-Fin (Chen et al., 2018) Bayesian lexicons (Eisenstein, 2017) Loughran-McDonald lexicon (Loughran and McDonald, 2011)
LIPI (Ghosh and Naskar, 2022)	sbert-chinese-qmc-finance[1] FinBERT (Araci, 2019)	Linear regression MLP
DCU_ML (Lyu et al., 2022)	BERT-Chinese (Devlin et al., 2019)	BERT-Senti (Proposed)
UOA (Zou et al., 2022b)	Bert-Base-Chinese (Devlin et al., 2019) RoBERTa-wwm-ext (Cui et al., 2021)	Astock (Zou et al., 2022a)
aiML (Qin et al., 2022)	FinBERT-tone (Yang et al., 2020)	Sec-Bert-Shape (Loukas et al., 2022) Astock (Zou et al., 2022a)
Yet (Zhuang and Ren, 2022)	FinBERT-Chinese [2] Mengzi-Fin (Zhang et al., 2021) RoBERTa-large-pair (Xu et al., 2020) RoBERTa-wwm (Cui et al., 2021)	Stochastic Weight Averaging MADGRAD Optimizer multi-sample dropout Modified-RoBERTa-wwm (Proposed)
UCCNLP (Trust et al., 2022)	SBERT (Reimers and Gurevych, 2019)	DPP-VAE (Proposed)
Jetsons (Gon et al., 2022)	Chinese-BERT (Devlin et al., 2019) xlm-roberta-large (Conneau et al., 2020)	Part of Speech

Table 1: Methods

$$MPP_{bearish} = (O_{t+1} - \min(L_{(t+1,T)}))/O_{t+1} \quad (3)$$

$$ML_{bearish} = (O_{t+1} - \max(H_{(t+1,T)}))/O_{t+1}, \quad (4)$$

where O_t and $H_{(t,T)}$ denote the opening price of day t and a list of the highest prices of day t to day T, respectively, and $L_{(t,T)}$ denotes a list of the lowest prices of day t to day T.

Based on the above labels, there are two task settings in ERAI shared task:

1. **Pairwise Comparison**: In the pairwise setting, there are two given opinions with MPP and ML labels. Models are asked to determine (i) whether the given opinion 1 will lead to higher MPP than the given opinion 2 and (ii) whether the given opinion 1 will lead to more loss than the given opinion 2. Thus, both would be binary classification tasks. We will use accuracy to evaluate the performances.

2. **Unsupervised Ranking**: In the unsupervised ranking setting, a pool of investors' opinions will be given, and the participants need to rank them with unsupervised methods. The goal is to find out the top 10% of posts that will lead to higher MPP. We will use the average MPP of the selected posts as the evaluation metrics.

2.2 Dataset Construction and Statistics

The dataset for the pairwise comparison setting is collected from Mobile01.[3] We manually checked the sentiment (bullish/bearish) in each opinion, and calculated MPP and ML based on the above equations. We labeled 574 posts (287 pairs), and further used 200 pairs as the training set and 87 pairs as the test set. The dataset for the unsupervised ranking setting is collected from PTT.[4] We also checked the sentiment (bullish/bearish) in each opinion manually and further obtained the MPP and ML labels. It is worth noting that, there are some posts that do not provide investment suggestions, but also follow the same template and are posted on the same platform as those that contain suggestions. We remain these posts in the pool to keep the dataset close to the real-world scenario. Thus, the posts that do not contain investment suggestions will get "nan" when annotating MPP and ML. Finally, a total of 210 posts are left in this set.

The original data for both tasks are written in Chinese. We use Google Translate API to prepare the English version. Participants can explore these tasks with the original data, translated data, or both.

3 Participants' Methods

Table 1 summarizes the methods used in this shared task. Both generation and classification language models are explored. Different kinds of domain-specific language models are also probed. Several lexicons are used for enhancing the performances, and some state-of-the-art architectures are used in the experiments. Tailor-made architectures and methods are also proposed by some teams.

[3] https://www.mobile01.com/

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

Team	MPP	Team	ML
Jetsons_1	62.07%	DCU-ML_1	59.77%
Yet_1	57.47%	DCU-ML_3	59.77%
Yet_2	57.47%	PromptShots_2	54.02%
Yet_3	57.47%	uoa_1	54.02%
LIPI_2	57.47%	aimi_1	52.87%
LIPI_1	54.02%	LIPI_2	50.57%
fiona	54.02%	fiona	48.28%
DCU-ML_1	52.87%	LIPI_3	48.28%
DCU-ML_3	52.87%	DCU-ML_2	45.98%
uoa_1	51.72%	PromptShots_1	45.98%
DCU-ML_2	51.72%	LIPI_1	44.83%
Jetsons_3	49.43%	Jetsons_2	41.38%
aimi_1	48.28%	PromptShots_3	41.38%
PromptShots_2	48.28%	Yet_1	40.23%
Jetsons_2	47.13%	Yet_2	40.23%
PromptShots_3	47.13%	Yet_3	40.23%
PromptShots_1	47.13%	Jetsons_1	37.93%
LIPI_3	44.83%	Jetsons_3	36.78%

Table 2: Pairwise Results (Accuracy).

Wiriyathammabhum (2022) prompt models for answering the instances in pair-wise setting, and aggregate lexicons' scores for unsupervised setting. Ghosh and Naskar (2022) ensemble the output of five models for both subtasks. Lyu et al. (2022) propose BERT-Senti, which is based on the notion that posts with more positive (negative) sentiment would lead to higher (lower) MPP. Both Zou et al. (2022b) and Qin et al. (2022) show that the method, AStock, talior-made for stock movement prediction cannot outperform vanilla pretrained language models in pairwise dataset. However, in unsupervised dataset, AStock outperforms vanilla pretrained language models. Zhuang and Ren (2022) explore different techniques such as the strategies of optimizer and drop out. Trust et al. (2022) propose DPP-VAE, and take the diversity and representation of the given opinion into consideration. Gon et al. (2022) provide a comparison of using various cross-lingual combination in training and testing.

4 Participants' Results

Table 2 and Table 3 show the results of participants' methods. It is worth noting that general language models perform better than domain-specific language models. For example, BERT-Chinese performs the best (Jetsons_1) in MPP comparison task, and Modified-RoBERTa-wwm (Yet_1,2,3) also performs well. However, both of them perform worse in ML comparison task. Additionally, positive/negative sentiment seems more related to

Team	Top 10% MPP	Team	Top 10% ML
PromptShots_2	24.39%	Baseline (Chen et al., 2021)	-2.46%
PromptShots_3	23.76%	Yet_3	-3.24%
PromptShots_1	22.53%	LIPI_1	-4.11%
LIPI_2	18.27%	aimi_1	-4.17%
Baseline (Chen et al., 2021)	17.61%	Yet_1	-4.35%
LIPI_1	17.46%	LIPI_3	-5.56%
UCCNLP_3	14.81%	Yet_2	-5.77%
Yet_3	14.61%	UCCNLP_3	-5.85%
aimi_1	14.02%	UCCNLP_1	-6.22%
DCU-ML_1	13.97%	UCCNLP_2	-6.77%
UoA_1	12.35%	PromptShots_1	-7.80%
Yet_2	12.10%	LIPI_2	-7.81%
LIPI_3	11.83%	DCU-ML_1	-8.25%
UCCNLP_2	11.34%	UoA_1	-9.39%
UCCNLP_1	11.10%	PromptShots_3	-12.33%
Yet_1	8.52%	PromptShots_2	-13.04%

Table 3: Unsupervised Results.

ML instead of MPP (DCU-ML_1). In the unsupervised setting, sentiment lexicons still play important roles (PromptShots_1,2,3). Most supervised results with the model trained with pair-wise setting dataset cannot outperform lexicon-based method and the baseline (Chen et al., 2021), which count the expert-like sentences in the post. On the other hand, the ML results in unsupervised setting imply that expert-like sentences matters in sorting out the opinions containing lower risk.

5 Future Directions

We want to highlight that before we try to sort opinions, we may need to first filter out those posts that do not contain trading ideas. For example, there are 57 of these kinds of posts in the unsupervised set. These posts follow the same format but may just ask questions. There are two reasons why we need to remove such posts. Firstly, in most cases, the models' input length is limited. Under this limitation, ideally, we should only use those considered important. Secondly, since this kind of posts does not contain opinion, putting them into a model may lead to incorrect claims and increase the noise. Following this line of thought, one of the future directions is to filter out both irrelevant and low MPP posts in the preprocessing process. On the other hand, the proposed idea can also use in a recommendation system for investors. Instead of only suggesting the relevant opinions as previous work (Liou et al., 2021), we think that recommending high potential suggestions would be more preferred in the investment scenario.

6 Conclusion

This paper introduces the methods explored in the ERAI shared task, and summarizes the performances of these methods. We think this is a pilot exploration for evaluating the rationales of

investors, and plan to dig into this direction more deeply in the future. The first step is exploring the role of argument in these tasks. We will present several datasets for extracting argument features from financial opinions, and we think that it will be useful in scoring investors' opinions. The enlarged dataset for evaluating investors' opinions will also be proposed. Please refer to the FinArg@NTCIR for more details.[5]

Acknowledgments

This paper is based on results obtained from a project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). This research was also partially supported by National Science and Technology Council, Taiwan, under grants MOST 110-2221-E-002-128-MY3 and MOST 110-2634-F-002-050-.

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. 2021. Evaluating the rationales of amateur investors. In *Proceedings of the Web Conference 2021*, pages 3987–3998.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. 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.

[5] http://finarg.nlpfin.com/

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.

Jacob Eisenstein. 2017. Unsupervised learning for lexicon-based classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. 2016. Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 549–554, Berlin, Germany. Association for Computational Linguistics.

Sohom Ghosh and Kumar Sudip Naskar. 2022. Lipi at the FinNLP-2022 erai task: Ensembling sentence transformers for assessing maximum possible profit and loss from online financial posts. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alolika Gon, Sihan Zha, Sai Krishna Rallabandi, Parag Pravin Dakle, and Preethi Raghavan. 2022. Jetsons at the FinNLP-2022 erai task: Bert-chinese for mining high mpp posts. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Md Kamrul Hasan, James Spann, Masum Hasan, Md Saiful Islam, Kurtis Haut, Rada Mihalcea, and Ehsan Hoque. 2021. Hitting your MARQ: Multimodal ARgument quality assessment in long debate video. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6387–6397, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yi-Ting Liou, Chung-Chi Chen, Tsun-Hsien Tang, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. Finsense: an assistant system for financial journalists and investors. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 882–885.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65.

Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. 2022. FiNER: Financial numeric entity recognition for

XBRL tagging. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4419–4431, Dublin, Ireland. Association for Computational Linguistics.

Chenyang Lyu, Tianbo Ji, and Liting Zhou. 2022. Dcuml at the FinNLP-2022 erai task: Investigating the transferability of sentiment analysis data for evaluating rationales of investors. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Zhaoxuan Qin, Jinan Zou, Qiaoyang Luo, Haiyao Cao, and Yang Jiao. 2022. aiML at the FinNLP-2022 erai task: Combining classification and regression tasks for financial opinion mining. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Gabriella Skitalinskaya, Jonas Klaff, and Henning Wachsmuth. 2021. Learning from revisions: Quality assessment of claims in argumentation at scale. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1718–1729, Online. Association for Computational Linguistics.

Paul Trust, Rosane Minghim, Ahmed Zahran, and Evangelos Milos. 2022. Uccnlp at the FinNLP-2022 erai task: Determinantal point processes and variational auto-encoders for identifying high-quality opinions from a pool of social media posts. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Peratham Wiriyathammabhum. 2022. Promptshots at FinNLP-2022 erai task: Pairwise comparison and unsupervised ranking. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Frank Xing, Lorenzo Malandri, Yue Zhang, and Erik Cambria. 2020. Financial sentiment analysis: An investigation into common mistakes and silver bullets. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 978–987, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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*.

Yan Zhuang and Fuji Ren. 2022. Yet at the FinNLP-2022 erai task: Modified models for 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.

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.

Jinan Zou, Haiyao Cao, Lingqiao Liu, Yuhao Lin, Ehsan Abbasnejad, and Javen Qinfeng Shi. 2022a. Astock: A new dataset and automated stock trading based on stock-specific news analyzing model. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing*.

Jinan Zou, Haiyao Cao, Yanxi Liu, Lingqiao Liu, Ehsan Abbasnejad, and Javen Qinfeng Shi. 2022b. Uoa at the FinNLP-2022 erai task: Leveraging the label information for financial opinion mining. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.