UOA at the FinNLP-2022 ERAI Task: Leveraging the Class Label Description for Financial Opinion Mining

Jinan Zou, Haiyao Cao, Yanxi Liu, Lingqiao Liu, Ehsan Abbasnejad, Javen Qinfeng Shi
Australian Institute for Machine Learning, The University of Adelaide
jinan.zou, haiyao.cao, yanxi.liu@adelaide.edu.au
lingqiao.liu, ehsan.abbasnejad, javen.shi@adelaide.edu.au

Abstract

Evaluating the Rationales of Amateur Investors (ERAI) is a task about mining expert-like viewpoints from social media. This paper summarizes our solutions to the ERAI shared task, which is co-located with the FinNLP workshop at EMNLP 2022. There are 2 sub-tasks in ERAI. Sub-task 1 is a pair-wised comparison task, where we propose a BERT-based pre-trained model projecting opinion pairs in a common space for classification. Sub-task 2 is an unsupervised learning task ranking the opinions' maximal potential profit (MPP) and maximal loss (ML), where our model leverages the regression method and multi-layer perceptron to rank the MPP and ML values. The proposed approaches achieve competitive accuracy of 54.02% on ML Accuracy and 51.72% on MPP Accuracy for pairwise tasks, also 12.35% and -9.39% regression unsupervised ranking task for MPP and ML.

1 Introduction

Using textual information to guide investment decisions is not a novel topic in either financial or fintech settings. Many researchers have devoted endeavors to social media posts and tried to dig out the rationale underlying the standpoints. However, these works struggle to cope with a considerable amount of data in the information explosion era, which brings an unnecessary expense to computation efficiency. Moreover, posts with high rationality have more probability of leading to profitable outcomes than those less rational. Thus, selecting high-quality analytical opinions can be a meaningful first step in investment opinion mining.

The ERAI shared task (Chen et al., 2022) proposes the rationale evaluation challenge with the goal of mining opinions leading to higher maximal potential profit (MPP) and lower maximal loss (ML). This challenge uses forecasting skills as a proxy and focuses on amateur investors' viewpoints. Two settings are involved in this challenge, including 1) Pairwise Comparison, which aims to find posts with more rationality; 2) Unsupervised Ranking, which aims to sort out the posts leading to the highest MPP and lowest ML. Several related works have launched good pilots for high-quality mining reviews. The BERT model proposed by Devlin et al. (2019) has been proven efficient in many NLP tasks since it was published. Chen et al. (2021c) presented and summarized the opinion mining methods. Chen et al. (2021a) provides methods to measure forecasting skills from the text. Chen et al. (2021b) creatively introduces the MPP and ML values to support digging into the review quality. Moreover, their proposed dataset, which is utilized in this paper, is the first dataset focusing on revealing text rationals.

Our model is based on a pre-trained language model, and for the binary classification task, we propose a method that utilizes the class-label information, and then we fine-tuned BERT for the regression task. The official results show that our models achieve competitive performance on both tasks, indicating our approaches' effectiveness. We introduce the tasks and present our work as follows. Section 2 elaborates on the shared task ERAI and the datasets for sub-tasks Pairwise Comparison and Unsupervised Ranking. We introduce our methodology and models in Section 3 and present the experimental setup and official results in Section 4. Finally, we conclude our work in Section 5.

2 Shared tasks

The ERAI shared tasks aim to spark interest from NLP and financial communities and to launch a novel pilot with the perspective of text rationality evaluation. The shared tasks have two sub-tasks focusing on digging into investors' posts and sorting out those with higher possibilities leading to MPP and ML.

Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP), pages 122 - 126

2.1 Sub-task 1: ERAI-pairwise

In the pairwise comparison setting, models are asked to determine rational-amateur post pairs' MPP and ML labels. Each pair gives two opinion posts together with their MPP and ML values. Also, the model is asked to predict: 1) the MPP label based on whether post1 has higher MPP than post2; 2) the ML label based on whether post1 has lower ML than post2. According to the findings of Chen et al. (2021b), a rational post may lead to higher MPP and lower ML values.

2.2 Sub-task 2: ERAI-unsupervised

In the unsupervised ranking setting, models are asked to rank the investors' posts within an opinion pool by the MPP and ML values. Unsupervised models would be utilized in this sub-task where the given data only contains the posts without any other supplementary information. The ranked top 10% posts should be the group having the highest average MPP value or lowest average ML value.

3 Methodology

3.1 Sub-task 1: Binary Classification

Label information is essential for humans to accurately interpret the meaning of a limited number of training samples. We proposed a method that utilized the class-label information for the two given opinions. We use BERT (Devlin et al., 2019) as the Pre-trained Language Model(PLM) unless specified otherwise. Specifically, we consider the following process to project two opinions in a common space in order to classify the class using [CLS] token. We append the corresponding class name and a [SEP] token after each training opinion to implement the binary classification tasks (i.e.[CLS] opinion1 [SEP] opinion2 [SEP] MPP Label Info [SEP] ML Label Info [SEP], where MPP Label Info could be 'higher maximum possible profit' or 'lower maximum possible profit', and ML Label Info could be 'higher maximum loss' or 'lower maximum loss'). We took the representation of [CLS] token at the model's last layer and added a linear layer for outputting MPP and ML binary classification results in Figure 1. In this binary classification task, we use Binary Cross Entropy Loss (BCE loss) as the loss function, which reflects the distributions divergence between labels and predictions. The smaller the value of cross-entropy is, the closer the two probability distributions are. BCE

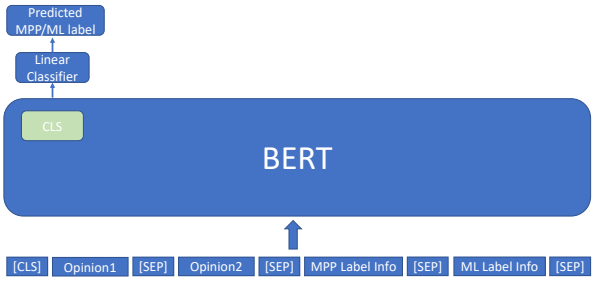

Figure 1: Overview of binary classification for sub-task 1 by leveraging the label information

loss can be described as equation (1):

$$\ell_{BCE} = -(y_i \cdot \log(\hat{y}_i) + (1-y_i) \cdot \log(1-\hat{y}_i)) \quad (1)$$

where $\hat{y}_i$ represents the predictions and y_i represents the labels.

3.2 Sub-task 2: Regression for the Unsupervised Ranking

In sub-task 2, the results are the ordered posts by the descending MPP and ascending ML, respectively. We fine-tuned a BERT model to adjust the regression task, whose outputs are ML and MPP values. We apply a dense pooling layer with dropout on the [CLS] embedding for the regression in sub-task 2 rather than just a dense linear layer in sub-task 1.

Mean squared error (MSE) loss is used to reflect the true error of the model in sub-task 2. The gradient of MSE loss increases as the loss increases and decreases as the loss tends to zero. The advantage of MSE in this task is that it converges effectively even with a fixed learning rate. MSE loss is as shown in the following equation (2):

$$\ell_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (2)$$

4 Experimental Setup and Evaluation

4.1 Dataset

The shared ERAI tasks aim to sort out the posts leading to higher MPP and lower ML. Regarding sub-task 1, the labeled and unlabeled datasets contain 200 and 87 pairs of posts, respectively. Each piece of the data consists of two posts, two MPP values with the MPP label, and two ML values with the ML label. The MPP label is determined by **Label "1": "MPP1" > "MPP2"; Label "0": "MPP1" < "MPP2"**. While the ML label relies

on **Label "1": "ML1" < "ML2"; Label "0": "ML1" > "ML2"**. This sub-task is asked to determine the MPP and ML labels of the post pairs in the unlabeled dataset. As Figure 2 shows, the labeled dataset containing 200 posts has a relatively even data distribution (i.e. MPP label 1/ label 0 is 109/91 and ML label 1/ label 0 is 105/95). We use the same labeled dataset in both sub-task 1 and sub-task 2.

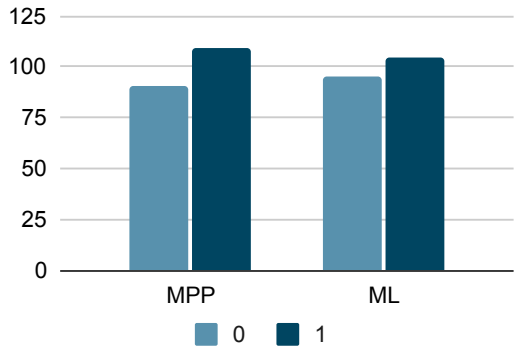

Figure 2: Distribution of MPP and ML label in labeled dataset

In terms of sub-task 2, the dataset contains 210 pieces of posts. This sub-task calls for an unsupervised model to dig into the posts' rationality and sort out the top 10% posts by the MPP and ML values, respectively.

4.2 Evaluation Metrics

According to the criteria of the ERAI challenge (Chen et al., 2021a), we use different evaluation methods for the two sub-tasks. We split 70% of the labeled dataset as training set and 30% as the validation set. In terms of sub-task 1, we use the accuracy to evaluate the model where the result indicates the model performance on two binary classifications (i.e., MPP label and ML label). We show the evaluation metric as the formula (3) (Linhares Pontes et al., 2022):

$$Accuracy = \frac{1}{n_{pair}} \sum_{i=1}^{n_{pair}} 1(\hat{y}_i = y_i) \qquad (3)$$

where $\hat{y}_i$ is the predicted label with the ground truth label y_i.

In sub-task 2, we use the average MPP value of the sorted top 10% to evaluate the model where a higher average MPP refers to better model performance. The evaluation metric shows the following formula (4):

$$Average = \frac{1}{n_{top}} \sum_{i=1}^{n_{top}} MPP_i \qquad (4)$$

where MPP_i represents the MPP value of the i_{th} post in the final rank list.

4.3 Hyperparameter setting

The models were trained on one Nvidia 2080Ti. The models were trained for 30 epochs with runtime ranging from 35 minutes to 1 hours. We used AdamW (Kingma and Ba, 2014) to optimize our model, and a learning rate of $2e-5$. The batch size is 8.

4.4 Experimental Evaluation

For optimizing purposes, we compared three pretrained models, including BERT-Base-Chinese (Wolf et al., 2020), a Chinese RoBERTa model named RoBERTa-wwm-ext (Cui et al., 2021), and a Chinese BERT-based model named Astock (Zou et al., 2022) that has been performed domain adaption by training the model with Masked-Language Model (MLM) loss on financial news articles.

Sub-task 1 MPP and ML Accuracy Value		
PLMs	MPP	ML
RoBERTa-wwm-ext (Cui et al., 2021)	**62.50%**	**57.50%**
Astock (Zou et al., 2022)	55.00%	52.50%
BERT-base-Chinese (Wolf et al., 2020)	60.00%	52.50%

Table 1: Experimental results for pairwise comparison in our split evaluation dataset

Sub-task 2 Average MPP and ML Value				
PLMs	Golden MPP	Golden ML	Pred MPP	Pred ML
RoBERTa-wwm-ext (Cui et al., 2021)	6.51%	-10.92%	3.2%	-3.21%
Astock (Zou et al., 2022)	**9.36%**	**-10.58%**	**4.23%**	**-3.11%**
BERT-base-Chinese (Wolf et al., 2020)	6.51%	-10.92%	2.86%	-3.85%

Table 2: Experimental results for the unsupervised ranking task in our split evaluation dataset, 'Golden' represents the real value and 'Pred' represents the predicted value

RoBERTa-wwm-ext achieved the best performance in MPP and ML accuracy on sub-task 1 as shown in Table 1. In sub-task 2, as shown in Table 2, the predicted MPP values and ML values of Astock are closer to the real values than other models, Golden MPP values are also approximately 3%

higher than others. Astock achieved outstanding performance than other PLM models on sub-task 2 in our split evaluation dataset. Therefore, we employed RoBERTa-wwm-ext for sub-task 1 and Astock for sub-task 2 due to the excellent performance as our final submission.

4.5 Official Released Results

The official results of each model across all teams are shown in Table 3. The listed MPP and ML results range from 62.07% to 44.83%, and 59.77% to 36.78%, respectively. Our result with 52.87% is ranked 2nd position (in Table 4) when taking the average of MPP and ML accuracy, which shows our model's high robustness and effectiveness. Specifically, UOA_1 yields an outstanding performance in MPP with an accuracy of 51.72%, and the accuracy of ML is 54.02%. Average MPP value and average ML values are used to evaluate the model performance in sub-task 2. Following the task instruction, a higher average MPP and a lower ML value suggest a better performance. Compared to the baseline (Table 5), UOA_1 provides an average MPP value of 12.35%, which is 5.26% lower than the baseline result. Regarding the average ML, the average value provided by UOA_1 is -9.39% lower than the baseline by 6.93%.

Accuracy			
Model Name	MPP	Model Name	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 3: Official results for pairwise comparison task

Average Accuracy	
Team Name	MPP+ML
DCU-ML	56.32%
UOA	**52.87%**
fiona	51.15%
PromptShots	51.15%
aimi	50.58%
Jetsons	50%
LIPI	49.43
Yet	48.85

Table 4: Best average accuracy on MPP and ML for each group

Pairwise sub-task 2 Averaged Value		
	MPP	ML
Baseline	17.61%	-2.46%
UOA-1	12.35%	-9.39%

Table 5: Official results for the unsupervised ranking task

In terms of model improvement, there are two directions we can move on. 1) Different layers of BERT capture different levels of semantic and syntactic information. The current UOA_1 model only uses the extracted features from the last layer, which loses much information. Future work can address this by fine-tuning the output features of each layer of the BERT model and invoking methods such as ablation strategies to extract more useful information from these features (Wang and Neumann, 2018). 2) A more considerable amount of data is preferred as BERT usually requires large quantities of data in regression tasks for a better result. Utilizing data augmentation techniques such as GPT-3 (Brown et al., 2020) could be a promising method.

5 Conclusion

This work presents the UOA team with how to tackle the ERAI shared tasks. For sub-task 1, we proposed a model by appending the class-label description from a pre-trained language model to accomplish the classification task. This suggests that our model is able to learn more discriminative features. Specifically, in sub-task 1, our proposed system achieved the second position considering the average of MPP and ML accuracy by statistical manually. For sub-task 2, we leveraged a regression framework to rank ML and MPP values. The official results show that our approaches could effectively solve the two tasks. Our models are simple but effective, and we achieved competitive performance on the shared tasks.

6 Limitations

Since our framework relies on a pre-trained model based on BERT, we have not considered other pre-trained models like GPT-3 (Brown et al., 2020), and will be explored in the future.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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

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

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.

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.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Elvys Linhares Pontes, Mohamed Benjannet, Jose G Moreno, and Antoine Doucet. 2022. Using contextual sentence analysis models to recognize esg concepts. *arXiv e-prints*, pages arXiv–2207.

Weiyue Wang and Ulrich Neumann. 2018. Depth-aware cnn for rgb-d segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 135–150.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jinan Zou, Haiyao Cao, Lingqiao Liu, Yuhao Lin, Ehsan Abbasnejad, and Javen Qinfeng Shi. 2022. Astock: A New Dataset and Automated Stock Trading based on Stock-specific News Analyzing Model. *arXiv e-prints*, page arXiv:2206.06606.