aiML at the FinNLP-2022 ERAI Task: Combining Classification and Regression Tasks for Financial Opinion Mining

Zhaoxuan Qin, Jinan Zou, Qiaoyang Luo, Haiyao Cao, Yang Jiao
The University of Adelaide
zhaoxuan.qin, jinan.zou, qiaoyang.luo@adelaide.edu.au
haiyao.cao, yang.jiao@adelaide.edu.au

Abstract

Identifying posts of high financial quality from opinions is of extraordinary significance for investors. Hence, this paper focuses on evaluating the rationales of amateur investors (ERAI) in a shared task, and we present our solutions. The pairwise comparison task aims at extracting the post that will trigger higher MPP and ML values from pairs of posts. The goal of the unsupervised ranking task is to find the top 10% of posts with higher MPP and ML values. We initially model the shared task as text classification and regression problems. We then propose a multi-learning approach applied by financial domain pre-trained models and multiple linear classifiers for factor combinations to integrate better relationships and information between training data. The official results have proved that our method achieves 48.28% and 52.87% for MPP and ML accuracy on pairwise tasks, 14.02% and -4.17% regarding unsupervised ranking tasks for MPP and ML. Our source code is available[1].

1 Introduction

The fast-growing financial social media has become a mainstream information source for investors. They prefer to follow high quality viewpoints with persuasive rationales. However, browsing numerous and noisy posts is time-consuming and inefficient. Therefore, automatically identifying high quality posts in the financial field is vital. FinNLP workshop of EMNLP-2022 (Chen et al., 2022) publishes a shared task regarding the above problem focusing on evaluating the rationales of amateur investors. There are two sub-tasks: pairwise comparison and unsupervised ranking for online posts. The posts are all from the financial social platforms of Chinese. Regarding sub-task1, we are asked to select high financial quality posts from pairs of posts. Regarding sub-task2, all given posts are required to rank by their potential financial quality. As evaluation, it is difficult to assess the quality of a post directly. Therefore, we propose and utilize the maximum possible profit (MPP) and the maximum loss (ML) in a certain period (Chen et al., 2021a) as the evaluation metric of opinion quality.

Several recent findings evaluate user-generated social media content, such as posts, tweets, and blogs, by NLP technology. Chen et al. (2021b) explores using AI models for detecting financial fine-grained sentiment tendencies. They proved the importance of mining the premises and evaluating the rationales of a financial opinion. Moreover, Patel and Ezeife (2021) investigated that Bidirectional Encoder Representations from Transformers (Bert) (Devlin et al., 2018) is an advanced deep learning model for fine-grained aspect-based opinion mining on social media posts. Besides, many natural language processing (NLP) technologies have been widely used to analyze financial-related domain information, for instance, stock price prediction (Mehtab and Sen, 2019) and financial sentiment analysis (Sohangir et al., 2018).

In our work, we leverage several pre-trained language models (PLM) for two tasks, roughly described as (1) Posts classification: Comparing financial quality for two given posts. (2) Posts ranking: Ranking all the test posts by the financial quality and selecting the top 10% of posts. We propose two strategies which are financial domain pre-training and multi-task learning. We are mainly concerned with shared word embedding learning through extracting financial semantic information on a large financial corpus for the pre-training process. Financial embedding would give precise and helpful semantic information for models in order to settle downstream financial tasks. Also, we introduce a multi-task learning approach that combines regression and classification tasks. In addition, multi-task learning can make good use of

[1] Source code: https://github.com/Zhaoxuanqin/EMNLP-competition

Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP), pages 127 - 131
December 8, 2022 ©2022 Association for Computational Linguistics

the relationship between tasks. Specifically, we combine our models with multiple linear classifiers regarding both tasks and optimize the models by joint weighted loss calculation.

At last, we introduce our methodology and solve the task as follows. Section 2 elaborates on the ERAI shared task 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 the final Section 5.

2 Datasets

We conduct experiments on two different datasets corresponding to two sub-tasks.

2.1 ERAI-Dataset-Pairwise

ERAI-Dataset-pairwise train dataset comprises 200 pairs of Chinese posts and their English translation version. Each piece of data includes 10 rows: Chinese post1 text, Chinese post2 text, English post1 text, English post2 text, post1 MPP value, post2 MPP value, post1 ML value, post2 ML value, MPP label, and ML label. MPP and ML labels represent the comparison result of MPP and ML values. For the MPP and ML values comparison settings, MPP Label "1" : "MPP1" < "MPP2"; MPP Label "0": "MPP1" > "MPP2" , on the other hand, ML Label "1": "ML1" < "ML2"; ML Label "0": "ML1" > "ML2". (Chen et al., 2021a). ERAI-Dataset-pairwise test dataset has 87 pieces data and identical settings with the train dataset except without the MPP and ML labels.

2.2 ERAI-Dataset-Unsupervised

The ERAI-Dataset-unsupervised dataset contains 210 Chinese rational posts and their English translation version without actual MPP and ML values. Our target is to rank all the posts by their potential MPP values and ML values, then select the top 10% of posts that will lead to higher MPP and lower ML.

MPP	**labels 1**	**labels 0**
	109	97
ML	**labels 1**	**labels 0**
	105	91

Table 1: ERAI-Dataset-pairwise labels distribution

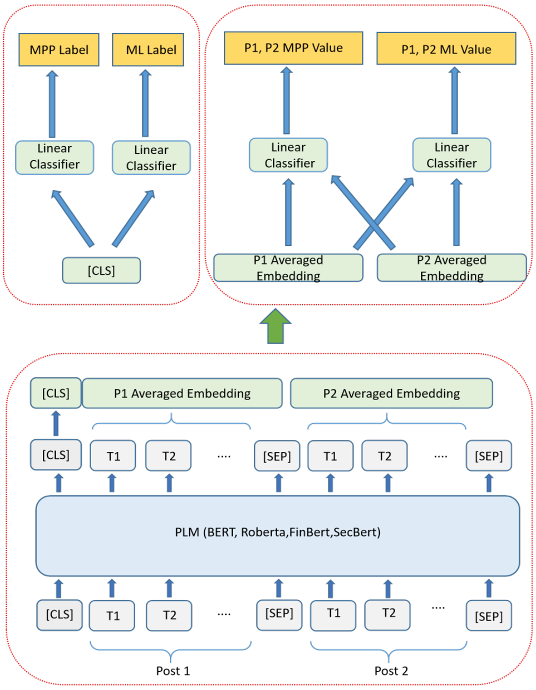

Figure 1: Our multi-learning process is composed of two modules. (a) text classification by the output of linear classifiers after [CLS] token from PLM. (b) text regression by the output of shared linear classifiers from the input post1 and post2 averaged sentence embedding. Two modules are trained together in the PLM, while they do not share the same linear classifiers in different tasks.

3 Method

3.1 In-domain Pre-training

In recent works, the pre-training of language models on specialized domains has been illustrated to have advantages for NLP downstream tasks. (Alsentzer et al., 2019; Yang et al., 2020). We compare three financial domain pre-training models to extract financial features that can improve the performance of the deep learning model.

Fin-Bert: We utilize Fin-Bert model (Yang et al., 2020), which is pre-trained on massive financial domain communication corpora including 4.9 billion tokens.

Sec-Bert-Shape: We also fin-tune Sec-Bert-Shape model (Loukas et al., 2022), which pre-trained on financial domain corpus on both tasks of our original dataset. The Sec-Bert-Shape model also trained word embedding by '[SHAPE]' pseudo-tokens. The English version datasets are applied to the models for extracting a financial representation

Astock: For Chinese models, We apply the Chinese financial domain adapted pre-trained RoBERTa model called Astock from Zou et al. (2022) on the Chinese financial news corpus .

3.2 Sub-task 1: ERAI Pairwise Comparison

Multi-task learning is explored for the task because the original dataset contains not only MPP and ML labels for text classification but specific MPP and ML values for text regression. Our proposed multi-task learning method is to classify the given post1 and post2 over MPP and ML labels. We consider this process a combination of text classification and a text regression task. Figure 1 illustrates the overall architecture of our multi-learning models. We add linear classifiers to process the [CLS] token embedding output from PLM, and shared linear classifiers to process the averaged post1 and post2 sentence embedding output from PLM. Specifically, the model is jointly optimized by the binary cross entropy loss and mean squared error text classification and regression. Furthermore, losses are weighted from those double tasks because there is a significant numerical difference, and it is essential to balance the losses. The formula 1 and formula 2 show how MSE loss and BCE loss are calculated where $\hat{y_i}$ represents the predicted labels and y_i represents the ground truth labels.

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

$$\ell_{bce} = -(y_i \cdot \log(\hat{y_i}) + (1 - y_i) \cdot \log(1 - \hat{y_i})) \quad (2)$$

The final loss we apply is composed of the weighted MSE loss and BCE loss, shown in the formula 3 below.

$$\ell_{total} = (1 - \omega) \cdot \ell_{bce} + \omega \cdot (\ell_{mse}^{mpp} + \ell_{mse}^{ml}) \quad (3)$$

As the total loss function (ℓ_{total}) defined, the learning loss function is modeled as the summation of weighted BCE loss (ℓ_{bce}) of classification and weighted MSE loss (ℓ_{mse}) for MPP and ML of regression. Specifically, The modeled MSE loss function consists of MPP and ML because the model can output MPP and ML values together in regression tasks. On the other hand, BCE loss is calculated by MPP and ML label together, so there is no need to sum the sub-BCE loss of MPP and ML.

3.3 ERAI Unsupervised Ranking

As sub-task2, there is no ground truth label for the posts in ERAI unsupervised dataset, so we use ERAI-Dataset-pairwise as the train and validation dataset on which we perform our experiment. This task requires a text regression task, and the model receives one sentence as input. Therefore, we separately trained the models which can output a single MPP or ML value. The PLM last hidden layer [CLS] token embedding is taken as the input of a one-dimension linear classifier to obtain predicted labels.

4 Experimental Setup and Evaluation

4.1 Evaluation Metric

All our experiments for both tasks use different evaluation metrics, where accuracy for the sub-task1 and averaged top 10% ranked values for the sub-task2 (Chen et al., 2021a). Accuracy determines how close the predicted labels are to their true labels:

$$Accuracy = \frac{1}{n_{sample}}\sum_{i=1}^{n_{sample}} 1(\hat{y_i} = y_i) \quad (4)$$

where $\hat{y_i}$ is the predicted values in our samples with their true labels y_i.

In sub-task2, we use the average MPP value of the sorted top 10% to evaluate the model where a higher average MPP or a lower ML refers to better model performance. The evaluation metric is shown in the formula 5:

$$Averaged\ Rank = \frac{1}{n_{top}}\sum_{i=1}^{n_{top}} y_i \quad (5)$$

where y_i represents i_{th} sample in the final MPP or ML rank list.

4.2 Experimental Details

We implement our approach with PyTorch 1.12.1 . We train all models for 30 epochs and choose the best model with the validation set. We use a batch size of 16, a maximum sequence length of 256, and a dropout probability of 0.1. For the optimizers, we utilize AdamW (Loshchilov and Hutter, 2017) with a learning rate of 2e-6.

4.3 Experimental Evaluation

We split the original ERAI pairwise train dataset into 80% for the new train dataset and 20% for the validation dataset. For both task1 and task2, we select four models which are Chinese base Bert (Devlin et al., 2018), Astock (Zou et al., 2022), Fin-Bert (Yang et al., 2020), and Sec-Bert-Shape (Loukas et al., 2022) respectively.

4.4 Experimental and Official results

We report the accuracy and mean rank based on five runs with different seeds for the above method. The averaged results of those 5 runs are shown in Table 2 and Table 3. As sub-task1, Table 2 has shown all the experimental performances on our validation dataset, which can be seen that Chinese Bert reaches a relatively high accuracy compared with the other three pre-trained models, which are 63.75% and 63% for MPP and ML prediction respectively. Astock and Sec-Bert-Shape perform best on MPP and ML values rank, respectively, with predicted averaged top 10% MPP 5.08% against true averaged top 10% MPP 9.04%, and predicted averaged top 10% ML -11.30% against true averaged top 10% ML -7.5%. We apply Chinese Bert, Astock, and Sec-Bert-Shape to the official test datasets.

Models	**MPP**	**ML**
Chinese-Bert	63.75%	63%
Fin-Bert	61%	54.20%
Sec-Bert-Shape	52.75%	59.30%
Astock	59.3%	60.25%

Table 2: MPP and ML accuracy

Models	**MPP**	**ML**
Chinese-Bert	3.94%	-12.05%
Fin-Bert	3.48%	-12.63%
Sec-Bert-Shape	3.30%	-11.30%
Astock	5.08%	-11.57%

Table 3: Average MPP and ML of Top 10% Posts

Our official results of submitted files are shown in Table 4. Our team achieves 48.28% and 52.87% regarding MPP accuracy and ML accuracy. As sub-task2, we report 14.02% and -4.17% over averaged top 10% MPP and ML values.

pairwise sub-task 1 accuracy		
team name	MPP	ML
aimi-1	48.28%	52.87%
pairwise sub-task 2 averaged values		
team name	MPP	ML
aimi-1	14.02%	-4.17%

Table 4: Official results

5 Conclusion

This paper describes a multi-task learning approach based on financial domain PLM for dealing with pairwise comparison and unsupervised ranking derived from rationales of amateur investors dataset. We demonstrate that joint loss optimization based on PLM can achieve competitive results. We also observed that Chinese-based PLM performs better than English-based PLM because the English translation cannot accurately express the exact meaning represented by the original Chinese version. For the results regarding both tasks, our models obtain an accuracy of 48.28% and 52.87% for MPP and ML labels in the first task. Besides, our second task achieves 14.02% and -4.17% for MPP and ML in the second task.

6 Limitations

There exist additional limitations in the current methods based on our method. Firstly, the Pre-trained Sec-Bert-Shape model tends to capture advanced representations of numerical tokens, while numerical token rarely appears in original datasets based on our observation. Secondly, we can not provide an efficient data augmentation method for a limited original dataset. The limitation of data may bring an overfitting problem for leading to an inferior result.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021a. 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. 2021b. A research agenda for financial opinion mining. *ICWSM*, pages 1059–1063.

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.

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*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. 2022. Finer: Financial numeric entity recognition for xbrl tagging. *arXiv preprint arXiv:2203.06482*.

Sidra Mehtab and Jaydip Sen. 2019. A robust predictive model for stock price prediction using deep learning and natural language processing. *arXiv preprint arXiv:1912.07700*.

Manil Patel and Christie I Ezeife. 2021. Bert-based multi-task learning for aspect-based opinion mining. In *International Conference on Database and Expert Systems Applications*, pages 192–204. Springer.

Sahar Sohangir, Dingding Wang, Anna Pomeranets, and Taghi M Khoshgoftaar. 2018. Big data: Deep learning for financial sentiment analysis. *Journal of Big Data*, 5(1):1–25.

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

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 preprint arXiv:2206.06606*.