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 Chen3
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):

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

(1)

\[ ML_{bullish} = (\min(L_{t+1,T}) - O_{t+1})/O_{t+1} \]

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

Table 1: Methods

\[
MPP_{\text{bearish}} = \frac{O_{t+1} - \min(L(t+1,T))}{O_{t+1}} \quad (3)
\]

\[
ML_{\text{bearish}} = \frac{O_{t+1} - \max(H(t+1,T))}{O_{t+1}} \quad (4)
\]

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
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, tailor-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 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.\textsuperscript{5}

References


Fourth Workshop on Financial Technology and Natural Language Processing, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.


