Jetsons at the FinNLP-2022 ERAI Task: BERT-Chinese for mining high MPP posts

Alolika Gon*, **Sihan Zha***, **SaiKrishna Rallabandi***, **Parag Pravin Dakle***, **Preethi Raghavan**
Fidelity Investments, AICoE, Boston
{alolika.gon, sihan.zha, saikrishna.rallabandi, paragpravin.dakle, preethi.raghavan}@fmr.com

Abstract

In this paper, we discuss the various approaches by the *Jetsons* team for the "Pairwise Comparison" sub-task of the ERAI shared task to compare financial opinions for profitability and loss. Our BERT-Chinese model considers a pair of opinions and predicts the one with a higher maximum potential profit (MPP) with 62.07% accuracy. We analyze the performance of our approaches on both the MPP and maximal loss (ML) problems and deeply dive into why BERT-Chinese outperforms other models.

1 Introduction

Natural language processing (NLP) has the potential to uncover meaningful insights from the vast amounts of unstructured data and impact the financial services industry. The use cases for financial NLP range from quantitative trading, portfolio selection, and risk assessment to speech recognition and customer chatbots on various unstructured sources, including transcripts of quarterly earnings calls, research reports, company filings, and social media chatter. People frequently express opinions about financial products, services, investments, and the stock market on social media. Such financial opinions can be effectively mined to provide recommendations and influence user/enterprise perception.

The Evaluating the Rationales of Amateur Investors (ERAI) shared task (Chen et al., 2022) focuses on opinions that would lead to profitable outcomes by using forecasting skills as a proxy. It is formulated as follows: given two opinions about a company extracted from Chinese social forums by amateur investors, predict the opinion that would lead to a higher profitable outcome or higher loss. We approach this problem using several strategies to represent and classify opinions, including: (1) using BERT-Chinese[1] on the original Chinese posts, (2) using RoBERTa (Liu et al., 2019) and other BERT (Devlin et al., 2018) variants on the English translated opinions, (3) using (2) in conjunction with POS tag features and (4) an ensemble of some of these approaches. Our approach using BERT-Chinese topped the test leaderboard for the MPP task.

We present the results of all these approaches and analyze how these models perform for the MPP task. We also examine what we may have lost in translation between Chinese and English and why the BERT-Chinese model outperforms the English-language BERT models.

2 Related Work

2.1 NLP on User generated content

The world of NLP has started focusing on user-generated content on the internet. There have been several works (Yadav et al., 2018; Wang et al., 2010) targeted at blogs, online forums (Yates et al., 2017), e-commerce platforms, and social media. Most of these works are oriented towards mining data from these sources, while of late, work targeted towards evaluating the opinion quality has garnered the community's interest. In (Diaz and Ng, 2018), authors present a survey in the context of e-commerce platforms. Chen et al. (2019) propose numeral attachment highlighting the relationship between cashtags and numerals in financial content. Lin et al. (2019) use sentiment on social media platforms to predict company sales while Xu and Cohen (2018) adopt tweets to predict stock movement. Basile et al. (2019) find that the style information of restaurant reviews can provide information about the authors. Zhang et al. (2019) show that authorship styles can predict the trafficker. Our current work follows the ideology of employing user-generated content online. Specifically, we are interested in comparing a pair of opinions presented by amateur investors and identifying the profitable

*These authors contributed equally to this work
[1]https://huggingface.co/bert-base-chinese

Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP), pages 141 - 146
December 8, 2022 ©2022 Association for Computational Linguistics

one among them.

2.2 Ranking Opinions

Feature-based approaches have been developed to rank argumentative comments (Wei et al., 2016) and product reviews (Eirinaki et al., 2012). In Ying and Duboue (2019), authors annotate a pilot dataset and classify rationales into four levels for educational purposes. In Chambers and Jurafsky (2008); Chambers et al. (2007), authors demonstrate how action words impact the narrative chain. In our work, we apply a similar strategy to opinions by grounding the model on action words in the opinion.

2.3 NLP based on Machine Translation

Several works have documented the advantages of employing a machine translation model to perform NLP tasks in a target language. Back translation has been a very useful part of several tasks such as sentence simplification (Vo et al., 2022), style transfer (Prabhumoye et al., 2018), semantic role labeling (Wu et al., 2022), etc. However, translation has been shown to cause confounding errors due to the errors in the translated content. In our current work, we highlight this by comparing the performance of models trained directly on Chinese and the models trained on the translated English version of the data. We identify two categories of translation errors and highlight them in our analysis.

3 Dataset and Methods

The training dataset contains 200 instances of opinion pairs in Chinese, their English translations, along with an MPP (maximum potential profit) label and an ML (maximum loss) label (Chen et al., 2021). In every instance, the pair of opinion posts have associated MPP_i and ML_i values where $i \in \{1, 2\}$, $\text{MPP}_i \in [0.0, 0.16)$ and $\text{ML}_i \in (-0.24, 0.0]$. If opinion 1 leads to higher MPP than opinion 2 i.e., $\text{MPP}_1 \geq \text{MPP}_2$, the MPP label is 1, otherwise 0. Similarly, if opinion 1 leads to higher loss than opinion 2 i.e. $\text{ML}_1 \leq \text{ML}_2$, the ML label is 1, otherwise 0. Out of the 200, there are only seven instances where the absolute difference between ML values of opinion 1 and 2 is greater than 0.1. Similarly, for MPP values, there are only six instances where the absolute difference between opinions 1 and 2 is greater than 0.1. The dataset distribution of ML and MPP labels as shown in Table 1. The test dataset contains 87 pairs of Chinese opinion posts and their English translations.

Dataset	Labels	**ML**	**MPP**
Training	1	105	109
	0	95	91
Testing	1	63	44
	0	24	43

Table 1: Distribution of labels

For the "Pairwise Comparison" subtask, given pairs of opinions in Chinese and their English translations as input, we train two separate classification models to predict the MPP label and ML label, respectively. We describe some of our approaches in the following subsections.

3.1 BERT-Chinese (BBC)

Since Chinese is the original language of the posts, we consider using a language model to process the information embedded in Chinese. We choose the 'bert-base-chinese' model (BBC), a pre-trained Chinese model based on the 'bert-base-uncased' model. We finetune a classification model based on the pre-trained BBC model by adding a binary classification layer on top of the pre-trained model. We tokenize and append the opinion pairs separated by a *[SEP]* token and feed it to our models as input. The learning rate is set to $1e-5$, and the model is trained for 20 epochs.

3.2 Using POS Tags and Named Entities

Given the small size of the training set, we consider hand-crafted features to train our classification models. We fine-tune 'xlm-roberta-large' (Conneau et al., 2019) on verbs (XRL-VERBS in table 2) and named entities (XRL-ENTITIES in table 2) extracted from the opinions using the 'spacy' python library [2]. Instead of feeding the entire posts as input to the models, we use space-separated verbs or named entities. The tokenization, input sequence, and final classification layer for both models are generated as described in subsection 3.1. The learning rate is set to $8e-6$, and the models are trained for ten epochs.

3.3 Ensemble

We also develop an ensemble model combining the Chinese posts and the corresponding English translations. We feed the Chinese posts into the

[2]https://spacy.io/usage/linguistic-features

Model	MPP-test	ML-test
BBC	**62.07**	37.93
XRL-VERBS	49.43	36.78
XRL-ENTITIES	53.49	59.30
ENSEMBLE	47.13	41.38

Table 2: Results of the experiments on the test set.

BBC model and the English posts into the 'xlm-roberta-large' model, respectively. We concatenate the final hidden states from the two models and add a linear layer on the combined hidden states to generate the binary classification results. Considering the complexity of the model, a dropout layer is added with a dropout ratio set to 0.3, and weighted cross-entropy loss is used as the loss function. The learning rate is set to $1e-5$, and the model is trained for 15 epochs.

4 Experimental Results

All models are trained using 10-fold cross-validation on the training set. The model corresponding to the best fold accuracy is used to obtain predictions on the test set. Table 2 shows the accuracy scores on the test set. The table shows that the *BBC* model performs the best on the test set for the MPP task with an accuracy of 62.07%. The *XRL-ENTITIES* model performs the best for the ML task with an accuracy of 59.30%.

5 Analysis

This section focuses on analyzing the impact of using the original Chinese posts for classification. The analysis is carried out in two ways - understanding the translation errors and probing the BBC model. All analysis presented in this section is for the MPP classification task.

5.1 Translation Errors

Out of the 87 test instances, the BBC model incorrectly classifies 33 instances. We further filter these instances using three steps. First, train an equivalent English model, M_e, for the MPP classification task. Second, Let S_e be the set of instances where the model M_e makes an incorrect classification. Lastly, filter S_e to obtain S_{e-c} by keeping instances that BBC correctly classified.

The BBC model is the 'bert-base-uncased' model further pre-trained on the Chinese Wikipedia data. Therefore, for M_e we use the 'bert-base-uncased' model and obtain S_{e-c} containing 17 instances. One annotator fluent in both languages manually analyzed the English translations of these 17 posts. The observed errors are divided into two categories:

1. **Literal translation of idioms** - In some cases, the Chinese text span that represents an idiom is translated literally and not contextually. For example, '盤中給賣掉，現在給我漲起∼∼氣死人' in the provided dataset is translated to 'Sold it on the plate, and now give me up ∼∼ Furious people' where the span '盤中給賣掉' literally translates to 'Sold it on the plate'. However, contextually, the span means 'sold it in the middle of the day'.
2. **Missing words/insertion of new words** - In some cases, the translation of a span of Chinese text does not match the actual meaning or inserts new words. For example, in the provided dataset, '發哥每天的利多還是比利空多 但股價磨人阿 三不五時還會破底 支撐都不是支撐 能抱的住真的很厲害' is translated to 'Big Brother's daily Lido is still Billy, But the stock price is grinding It will break the bottom of three or five o'clock Support is not support It's really amazing to hold it.' However, the span '每天的利多還是比利空多' actually translates to 'is more bullish than bearish every day.'

5.2 BBC Model Probing

The BBC model has been pre-trained first on English text, followed by Chinese text. Since the model has been trained in both languages, we evaluate the model using posts in different training and test languages. In addition to using the original English-translated posts, we also generate training and test datasets using Google Translate[3] to evaluate the effect of using another translation system. Table 3 shows the results of these experiments. The table reports the average test set accuracy across the ten folds and the best test set accuracy.

The results show that using Chinese as the training language and English as the testing language results in the highest accuracy and average accuracy. Additionally, we see an increase in the test set accuracy when the English posts generated using Google Translate are used, showing the impact of the translation errors. These observations lead to two questions - (1) if the BBC model vocabulary

[3]https://translate.google.com/

Train language	Test language	Avg. accuracy	Best accuracy
zh	zh	59.8	64
zh	en-old	63.9	66
zh	en-cor	**66.3**	**67**
en-cor	en-cor	62.3	66
en-cor	zh	47.6	52
zh	es	63.7	64

Table 3: Results of experiments using different languages for train and test set with the BBC model on the MPP label classification task. zh - Chinese, en - English, en-old - the original English posts, en-cor - the corrected English posts, es - Spanish.

is in Chinese, what information is the model extracting from English tokens to classify the posts correctly? and (2) what happens when a third language is used for the test set?

To answer the first question, we use the *transformers-interpret*[4] package to visualize the token level attentions to understand which tokens helped the model in correctly classifying the posts. For this experiment, we use two models - the BBC model trained on Chinese posts but tested on the corrected English posts (BBC_{CE}), and a BBC model trained on Chinese posts and tested on the Chinese posts (BBC_{CC}). We look at two examples: both models make a correct prediction, and only BBC_{CE} makes a correct prediction. Figures 1 and 2 show two examples, first, where the model makes a correct prediction for both languages, and second, where the model predicts correctly only for the English language. The attention weights (in green) in Figure 1 show that the models mostly attend to the same tokens when making the prediction. However, this is not the case for the second example in Figure 2, where the model attends to different tokens when given the Chinese posts as input. The attention scores also show that the model significantly attends to *UNK* tokens. We intend to investigate this observation as part of our future work.

In the final set of analyses, we experiment by using Spanish for the test set to evaluate if the model can transfer the learning to another language owing to its impressive performance when using English for testing. Table 3 shows that using Spanish results in the best test set accuracy of 64%. Empirically, this seems to match the accuracy obtained when

[4]https://github.com/cdpierse/transformers-interpret

Figure 1: Test example showing the BBC model with token attentions for English and Chinese language with correct predictions for both languages

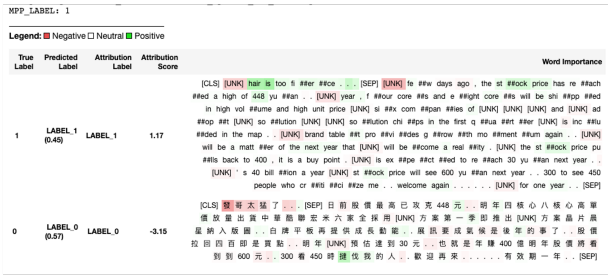

Figure 2: Test example showing the BBC model with token attentions for English and Chinese language with correct predictions for English

using Chinese for the test set. However, when analyzed, we observe that the model predicts class label 1 for all test samples resulting in high accuracy. This experiment yields two key observations - (1) the BBC model exhibits impressive performance on the English test set as it is pre-trained on the language, and (2) the accuracy metric cannot be used to evaluate models for this task owing to its class imbalance. Another metric, like Macro F1, can alleviate the class imbalance and help in better model comparison.

6 Conclusion

This paper discusses the models submitted for the ERAI Pairwise Comparison subtask organized at FinNLP 2022. Of the submitted models, the BERT-Chinese model trained on the Chinese posts ranks first on the MPP label leaderboard. We investigate why using Chinese posts over translated English posts results in higher accuracy and attribute the behavior to errors in translation. Additionally, we probe the BERT-Chinese model using different training and testing language combinations to evaluate the impact of two language pre-training. We show that the model did better when trained on Chinese posts and tested on English translations. Lastly, we show that the accuracy metric is not suited for the task owing to its inability to handle class imbalance.

References

Angelo Basile, Albert Gatt, and Malvina Nissim. 2019. You write like you eat: Stylistic variation as a predictor of social stratification. *arXiv preprint arXiv:1907.07265*.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797.

Nathanael Chambers, Shan Wang, and Dan Jurafsky. 2007. Classifying temporal relations between events. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 173–176.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2019. Numeral attachment with auxiliary tasks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1161–1164.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. 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, 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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Gerardo Ocampo Diaz and Vincent Ng. 2018. Modeling and prediction of online product review helpfulness: a survey. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 698–708.

Magdalini Eirinaki, Shamita Pisal, and Japinder Singh. 2012. Feature-based opinion mining and ranking. *Journal of Computer and System Sciences*, 78(4):1175–1184.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, Zihan Liu, Yan Xu, Cong Gao, and Pascale Fung. 2019. Learning to learn sales prediction with social media sentiment. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, pages 47–53.

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*.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000*.

Vy Vo, Weiqing Wang, and Wray Buntine. 2022. Unsupervised sentence simplification via dependency parsing. *arXiv preprint arXiv:2206.12261*.

Jia Wang, Qing Li, Yuanzhu Peter Chen, and Zhangxi Lin. 2010. Recommendation in internet forums and blogs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 257–265.

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200.

Han Wu, Haochen Tan, Kun Xu, Shuqi Liu, Lianwei Wu, and Linqi Song. 2022. Zero-shot cross-lingual conversational semantic role labeling. *arXiv preprint arXiv:2204.04914*.

Yumo Xu and Shay B Cohen. 2018. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1979.

Shweta Yadav, Asif Ekbal, Sriparna Saha, Pushpak Bhattacharyya, and Amit Sheth. 2018. Multi-task learning framework for mining crowd intelligence towards clinical treatment.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848*.

Annie Ying and Pablo Duboue. 2019. Rationale classification for educational trading platforms. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, pages 14–20.

Yiming Zhang, Yujie Fan, Wei Song, Shifu Hou, Yanfang Ye, Xin Li, Liang Zhao, Chuan Shi, Jiabin Wang, and Qi Xiong. 2019. Your style your identity: Leveraging writing and photography styles for drug trafficker identification in darknet markets over attributed heterogeneous information network. In *The World Wide Web Conference*, pages 3448–3454.

A Appendix

Table 4 shows the 10 fold cross validation accuracy scores for our best models in the MPP and ML subtask. The high variance in the scores is due to the dataset's small size.

k	MPP-test	ML-test
0	45	60
1	80	60
2	55	55
3	80	60
4	65	80
5	35	60
6	50	50
7	50	60
8	55	65
9	60	20

Table 4: Ten-fold cross validation accuracy scores of the *BBC* model for the MPP task and the *XRL-ENTITIES* model for ML task.