NTCIR 17 Conference: Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan
DOI: https://doi.org/10.20736/0002001277

WUST at the NTCIR-17 FinArg-1 Task

Mengjie Wu
School of Computer Science and Technology, Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology
Wuhan, China
1749565249@qq.com

Maofu Liu
School of Computer Science and Technology, Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology
Wuhan, China
liumaofu@wust.edu.cn

Tian Zhang
School of Computer Science and Technology, Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology
Wuhan, China
1634236713@qq.com

ABSTRACT

This article introduces how we deal with the FinArg-1 task of NTCIR17. In the FinArg-1 task, we have completed three subtasks which are argument classification, argument relation identification, and identifying relations in the social media dataset. In the experiments, we use the BERT model for the FinArg-1 three subtasks module.

KEYWORDS

Argument relation dentification, Argument unit detection, Argument classification

TEAM NAME

WUST

SUBTASKS

Argument Classification, Argument Unit Identification, and Argument Relation Identification.

1 INTRODUCTION

The WUST team participated in the NTCIR-17 FinArg-1 task [3]. This paper introduces the models and methods we used in this task and discusses the experimental results. Chen et al. [2] pointed out that Fintech had become a hot topic, allowing more academics to do natural language processing research from the financial domain. Argument mining is a popular study direction in natural language processing. In FinArg-1 task [1], there are three sub-tasks, namely Argument classification, argument relation identification and identifying relations in the social media dataset. In the argument classification subtask, we are asked to classify the given sentence into claim or premise. For the two remaining subtasks, we need to identify the relations (support/attack/none) of the given two sentences. But the corpus and language of these two subtasks are not the same. Our team regards this task as a text classification problem. We choose the BERT model for pre-training and classification and also added a kind of input text encoding. The rest of this report is organized as follows. Section 2 shows the related work of argument classification in financial domain. Section 3 introduces related models and methods. Section 4 shows the official experimental results and our analysis. Finally, some conclusions are drawn in Section 5.

2 RELATED WORK

When it comes to argument classification, we usually refer to the task of categorizing arguments in a text. Levy et al. [7] identify context-dependent claims (CDCs) by splitting the problem into smaller sub-problems. Rinott et al. [9] extend this work with a pipeline of feature-based models that find and rank supporting evidence from Wikipedia for the CDCs. However, neither of these approaches leverage the potential of word embeddings in capturing semantic relations between words.et al. Nils et al. [8] use both ELMo and BERT in the context of open domain argument search. For the first time, it was shown how the power of contextual word embeddings can be utilized to classify and cluster topic-related arguments.

When it comes to argument relation identification, we usually refer to the task of identifying the relationship between different arguments in a text. The goal of this task is to identify the relationship between two arguments, such as support, opposition, neutrality, etc. Yohan et al. [5] add four logical and theoretical mechanisms to make categorization better based on argument mining in categorizing argumentative relationships (support, attack, and neutrality) between statements.

3 METHODS

3.1 Pre-trained Language Models

BERT [4] is a bi-directional encoding representation model derived from the transformers model. The aim is to pre-train deep bi-directional representations from unlabeled text by computing conditionals that are common to both left and right contexts. Robustly Optimized BERT Pretraining Approach [10] (RoBERTa) is based on the improvement of BERT, which uses a dynamic mask. In this paper, we use the BERT model to do the NTCIR-17 FinArg-1 task. In the BERT model, we use text vector as the embedding. Taking into account the text length of the three tasks, we set the dimension of a batch as batchsize $\times 128 \times 768$.

For argument classification task 1, we consider it as a text binary classification problem. We directly use BERT model for text processing, BERT itself has a strong learning ability, the input text is transformed into the corresponding word vector. According to the known label training to get the results, later through softmax to compare and analyze the similarity between the text to determine whether it is premise or claim.

Both subtask 2 and subtask 3 tasks consist of determining how to identify the relations (support/attack/none) of the dataset. But subtask 2 is in English and subtask 3 is in Chinese. To determine the relationship between two texts, from a semantic point of view, it can be understood as to determine whether the relationship between two sentences is a cascade, cause and effect, or transitive relationship. Therefore, we use BERT as the pre-training model, we tried to use RoBERTa but the result is not as good as BERT. For subtask 1 and subtask 2, we use bert-base-uncased for training.Since the language of subtask 3 text is Chinese, we use bert-base-chinese for training.

We draw the entire model architecture as shown in Figure 1.

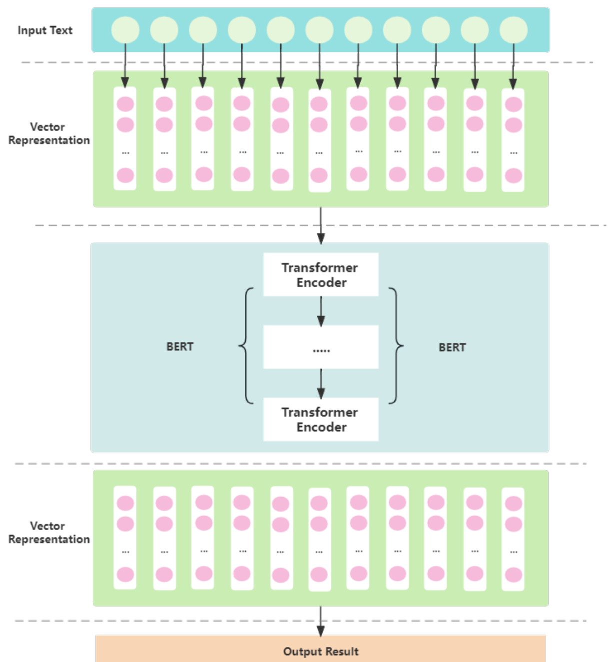

Figure 1: The model architecture diagram of BERT

4 EXPERIMENTS

4.1 Dataset and Evaluation Metrics

In this experiment, the dataset we use is provided by NTCIR-17. The organizers provide an analysis report from which we can derive a specific distribution of the number of the 3 subtasks datasets. In argument unit identification dataset,the quantity of training dataset, development dataset and test dataset are 7753, 969 and 969 respectively. In argument relation identification dataset, the quantity of training dataset, development dataset and test dataset are 5521, 690 and 690 respectively. In social media dataset, the quantity of training dataset, development dataset and test dataset are 6518, 815 and 815 respectively.

We use Adam [6] as the optimizer. We have fine-tuned the BERT model to better fit all three subtasks. In Task 1, we set the learning step to 0.00003, batchsize to 32, epoch to 2, and max_sequence length to 128. In Task 2, we set the learning step to 0.00003, batchsize to 16, epoch to 4, and input sequence length to 512. In Task 3, we set the learning step to 0.00003, batchsize to 8, and epoch to 3, input sequence length to 32. According to official evaluation criteria, we use the Micro-F1 and Macro-F1 score to evaluate the experimental results. The calculation process of Macro-F1 is to first calculate F1 of each category separately, and then average F1 of each category, with the same weight of each category. The formula of F1 is as follows:

Table 1: Results of argument unit identification

Team	Micro-F1	Macro-F1	Weight-F1
TMUNLP-1	76.57%	76.55%	76.59%
IDEA-1	76.47%	76.46%	76.48%
TUA1-1	76.37%	76.36%	76.38%
IMNTPU-2	76.06%	76.05%	76.07%
TMUNLP-3	76.06%	76.04%	76.07%
TMUNLP-2	75.95%	75.94%	75.97%
MONETECH-3	75.54%	75.53%	75.56%
IMNTPU-1	75.44%	75.31%	75.40%
MONETECH-1	75.13%	75.13%	75.12%
MONETECH-2	75.03%	75.02%	75.04%
TUA1-0	74.61%	74.56%	74.62%
WUST-1	**74.41%**	**74.41%**	**74.41%**
LIPI-3	73.89%	73.86%	73.90%
IDEA-3 (Late)	73.68%	73.68%	73.69%
LIPI-1	73.48%	73.47%	73.49%
LIPI-2	73.27%	73.27%	73.28%
SCUNLP-1-2	71.10%	71.07%	71.02%
SCUNLP-1-3	71.10%	70.53%	70.73%
SCUNLP-1-1	68.73%	68.62%	68.53%
WUST-2	69.04%	67.76%	68.07%
IMNTPU-3	56.97%	56.82%	56.70%

$$F1 = \frac{2 * (precision * recall)}{precision + recall} \quad (1)$$

$$precision = \frac{TP}{TP + FP} \quad (2)$$

$$recall = \frac{TP}{TP + FN} \quad (3)$$

4.2 Experimental results

Table 1, Table 2 and Table 3 show the results of the three subtasks of FinArg-1 respectively. Figures 2,3 and 4 show the data distribution of the training set of the three subtasks respectively. It can be clearly seen that the data distribution of the three tasks is uneven, especially argument relation identification task. Our results are WUST_1 and WUST_2.

In Table2, it shows that for all results, Micro-F1 is higher than Macro-F1 on the whole, because Macro-F1 adds the influence factor of uneven data distribution. From Table 3, we can see that Micro-F1 values are generally a bit higher than Macro-F1, but not as big as the difference in Table 2. As shown in Figures 2 and 3, in the social media dataset, the no relation category accounts for 11% of the data samples, but in the argument relation identification dataset, the no relation category accounts for only 1% of the data samples.

Table 2: Results of argument relation identification

Team	Micro-F1	Macro-F1	Weight-F1
TUA1-1	85.65%	61.50%	84.86%
LIPI-3	79.42%	60.22%	78.90%
TMUNLP-2	82.03%	57.90%	81.57%
TMUNLP-1	81.88%	57.36%	81.45%
TMUNLP-3	81.88%	56.72%	81.52%
TUA1-2	81.30%	56.26%	80.76%
TUA1-0	85.94%	55.36%	85.13%
SCUNLP-1-3	72.17%	54.06%	72.35%
WUST-1	**78.70%**	**53.97%**	**77.93%**
IMNTPU-2	82.61%	52.97%	82.14%
IDEA-3(Late)	81.74%	51.85%	80.88%
LIPI-1	80.72%	51.35%	80.09%
IDEA-1	80.58%	51.12%	79.89%
LIPI-2	80.29%	51.08%	79.79%
IMNTPU-3	80.72%	50.73%	79.67%
SCUNLP-1-2	68.55%	49.00%	68.57%
IMNTPU-1	78.99%	47.36%	76.54%
SCUNLP-1-1	68.70%	45.68%	68.05%
IDEA-2	57.10%	29.18%	59.39%

Table 3: Results on the social media dataset

Team	Micro-F1	Macro-F1	Weight-F1
Quack-2	71.66%	73.94%	71.35%
WUST-1	**70.55%**	**70.64%**	**70.30%**
Quack-1	67.85%	70.28%	67.30%
LIPI-3	64.79%	69.45%	64.09%
Quack-3	65.52%	66.88%	63.76%
SCUNLP-2-3	62.58%	66.39%	63.37%
SCUNLP-2-1	56.81%	59.76%	57.08%
SCUNLP-2-2	56.56%	59.6%	57.21%
LIPI-2	56.81%	58.28%	56.89%
LIPI-1	59.14%	57.30%	59.62%
CYUT-2	68.22%	49.62%	68.22%
TMUNLP-1	46.38%	35.37%	45.84%
IMNTPU-1	52.88%	34.77%	48.73%
TMUNLP-3	45.28%	32.48%	43.45%
TMUNLP-2	41.96%	31.69%	41.99%
IMNTPU-2	48.71%	24.64%	40.50%
CYUT-3	29.20%	23.45%	30.56%
CYUT-1	24.54%	20.94%	25.54%

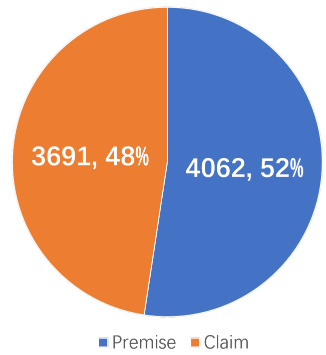

Figure 2: The data distributions of argument unit identification dataset

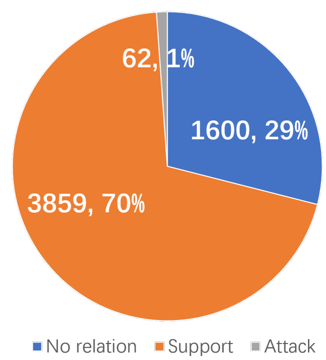

Figure 3: The data distributions of argument relation identification dataset

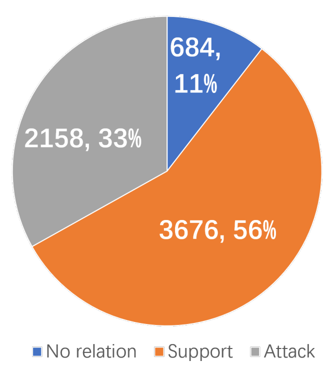

Figure 4: The data distributions of social media dataset

5 CONCLUSIONS

In this report, we employ the BERT model to do the three subtasks of FinArg-1. However, in Subtask 2, we did not have a solution that could solve the problem of the large difference between the two results Micro-F1 and Macro-F1 due to the uneven distribution of the dataset.

REFERENCES

[1] Alaa Alhamzeh, Romain Fonck, Erwan Versmée, Elöd Egyed-Zsigmond, Harald Kosch, and Lionel Brunie. 2022. It's Time to Reason: Annotating Argumentation Structures in Financial Earnings Calls: The FinArg Dataset. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 163–169. https://doi.org/10.18653/v1/2022.finnlp-1.22

[2] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. NLP in FinTech Applications: Past, Present and Future. *CoRR* abs/2005.01320 (2020). arXiv:2005.01320 https://arxiv.org/abs/2005.01320

[3] Chung-Chi Chen, Chin-Yi Lin, Chr-Jr Chiu, Hen-Hsen Huang, Alaa Alhamzeh, Yu-Lieh Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2023. Overview of the NTCIR-17 FinArg-1 Task: Fine-Grained Argument Understanding in Financial Analysis. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information*

Access Technologies. https://doi.org/10.20736/0002001323

[4] 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 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[5] Yohan Jo, Seojin Bang, Chris Reed, and Eduard Hovy. 2021. Classifying argumentative relations using logical mechanisms and argumentation schemes. *Transactions of the Association for Computational Linguistics* 9 (2021), 721–739.

[6] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[7] Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 1489–1500.

[8] Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. *arXiv preprint arXiv:1906.09821* (2019).

[9] Ruty Rinott, Lena Dankin, Carlos Alzate, Mitesh M Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence-an automatic method for context dependent evidence detection. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. 440–450.

[10] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A Robustly Optimized BERT Pre-training Approach with Post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*. Chinese Information Processing Society of China, Huhhot, China, 1218–1227. https://aclanthology.org/2021.ccl-1.108