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/10.20736/0002001288

TUA1 at NTCIR-17 FinArg-1 Task

Daichi Yamane
Tokushima University
Tokushima,Japan
c612335042@tokushima-u.ac.jp

Fei Ding
Tokushima University
Tokushima,Japan
c502147003@tokushima-u.ac.jp

Xin Kang*
Tokushima University
Tokushima,Japan
kang-xin@tokushima-u.ac.jp

ABSTRACT

This paper reports on the results produced by the TUA1 team in the Earnings Conference Call (ECC), Task 1 of Finarg-1 of NTCIR-17. The ECC is divided into two subtasks. One is Argument Unit Identification (AUI) and the other is Argument Relation Identification (ARI).

There are two proposed methods. The first is to tune a pre-trained model based on the transformer architecture using prompts. This method was applied to both Argument Unit Identification and Argument Relation Identification. The second approach employs Cost-Sensitive Learning on pre-trained models, which were previously tuned. This was exclusively used for Argument Relation Identification

In the provided training and validation data for Argument Relation Identification, the correct labels were markedly unbalanced, with some specific labels being notably scarce. Cost-Sensitive Learning proves effective for such unbalanced datasets, often yielding higher results than pure pre-trained models alone. In our experiments involving prompt tuning, we leveraged the Weighted Random Sampler technique to further enhance accuracy on the unbalanced data.

Experiments using the aforementioned methods revealed that we achieved the best results for Argument Relation Identification, and secured third place for Argument Unit Identification.

KEYWORDS

Transformer, Fine-Tuning, Argument Mining, Financial Engineering

TEAM NAME

TUA1

SUBTASKS

Argument Unit Identification (English)
Argument Relation Identification (English)

1 INTRODUCTION

With the advancement of natural language processing technology, the field of Artificial Intelligence (AI) has garnered significant public attention [1]. AI technology is being explored and utilized across various domains, with the financial sector being a notable example. Within finance, AI has found diverse applications, and the recent challenge tackled by the TUA1 team pertains to Argumentation Mining. This technique aims to elucidate the relationships between texts and has been an area of extensive research for some years. However, the application of Argumentation Mining within the financial realm remains nascent, presenting ample opportunities for development [2]. In this paper, we use prompt-based tuning of pre-trained models built on transformer and pre-trained models employing Cost-Sensitive Learning methods or Weighted Random Sampler methods to improve the F1-score of argument mining in this area. The goal of this paper is to improve the F1-score of argument mining in this area.

The TUA1 team participated in FinArg-1 of NTCIR-17. We took part in both the Argument Unit Identification subtask and the Argument Relation Identification subtask.

In the Argument Unit Identification subtask, the dataset's sentiment labels were divided into two categories: Premise and Claim. We conducted experiments using the F1-score as our evaluation metric. Initially, the pre-trained model was fine-tuned with training data. Subsequently, parameters were adjusted using validation data. The model then employed test data to predict sentiment labels, which were output as results. Also, all text in the distributed dataset is written in English.

In the Argument Relation Identification subtask, the sentiment labels in dataset are categorized as Attack, Support, and No detected. We utilize the F1 score as the primary evaluation metric for our experiments. Much like the previous subtask, we first tune the pre-trained model using the training data, then adjust parameters with the validation data. Subsequently, sentiment label predictions are made using the test data, and results are outputted. All datasets for this subtask consist of English text. It's noteworthy that there's a significant unbalance in the sentiment labels within these datasets. To address this, we experimented with the Weighted Random Sampler and Cost-Sensitive Learning methods. Both have been recognized for producing effective results in classification tasks involving unbalanced datasets [5].

In the Argument Unit Identification category, the TUA1 team submitted results from two experiments and secured a third-place ranking in F1-score among all participants. For Argument Relation Identification, the TUA1 team presented three sets of results and achieved the highest F1-score compared to all other participants [3].

This paper is organized as follows. Section 2 provides a detailed description of the Argument Unit Identification task and the proposed method used for it. Section 3 provides a detailed description of the Argument Relation Identification task and the proposed method used for it. Section 4 describes the experimental results and analyzes the results. Section 5 concludes and describes future prospects.

2 ARGUMENT UNIT IDENTIFICATION

The Argument Unit Identification subtask aims to classify the input text into two categories, Premise and Claim. An example dataset is shown below:

> Sentence 1 : I mean, sometimes it's not that you came up with some brilliant strategy, it's just like really good work consistently over a long period of time.
>
> Label : 0

We use the mapping $f_u : 0 \rightarrow$ "premise", $1 \rightarrow$ "claim" to map the corresponding labels.

We employed various pre-trained models for experimentation on this subtask. For example, ChatGLM with p-tuning [1], T5, BERT, FinBERT [2], and a T5 model fine-tuned with the financial phrasebank dataset [3]. Our experiments were repeated using these models. Evaluations were based on the F1-score, derived from the subsequent metrics.

$$\text{F1} = \frac{2 * \text{recall} * \text{precision}}{\text{recall} + \text{precision}}$$

In this subtask, there are three types of F1-score used as evaluation indicators: Micro F1-score, Macro F1-score, and Wighted F1-score. Each of them is explained below[4].

> Micro-F1 : By counting the sum of true positives (TP), false positives (FP) and false negatives (FN), we can calculate the metrics globally.
>
> Macro-F1: We can calculate the metrics for each label and find their unweighted mean. This does not take into account the imbalance of labels.
>
> Weighted-F1: Calculate the metrics for each label and determine their mean, weighted by support (the number of actual instances for each label). This modification of the "macro" metric accounts for the uneven label distribution, possibly resulting in an F-score that doesn't lie between precision and recall.

Finally, we used the pure T5 model and the T5 model with fine-tuning using the financial phrasebank dataset for the submission results.

Following this, we provide a detailed description of both the T5 model and the financial phrasebank dataset.

2.1 T5

The Text-to-Text Transfer Transformer (T5) is a model based on the Transformer architecture. T5 operates in a Text-to-Text format, producing text outputs in response to text inputs [4].

Prompt-based Learning and Instruction Fine-tuning have been verified to be effective in improving the embedding of inputs. For example, instead of implicitly providing certain parameters to the model, they are directly input in the form of questions. To address different Argument Analysis subtasks, we tested a large number of potential prompts and instructions, selecting a generic prefix from them. For each subtask, we maintain one long instruction and one short instruction, respectively. The prompts are listed in Table 1 and Table 2.

Subtasks	Prompt / Instruction
Short Instruction for AUI	Premise or Claim:
Long Instruction for AUI	###Instruction: Which view does the following argument belong to? Premise or Claim? ### Sentence: {text} ### Argument:

Short Instruction for ARI	Judge the relationship between the two sentences.
	Attack/Support/None: {text_1}
	{text_2}

Long Instruction for ARI	Below are two sentences that contain opinions. Please judge the logical relationship between sentence 1 and sentence 2. The relationship can only be among Attack, Support, or no-relation. ### Sentence 1: {text_1} ### Sentence 2: {text_2}

Table 1: Prompt and instruction for different subtasks. {text} and {label} stand for the original dataset inputs and outputs respectively.

Given that both input and output are always strings in the T5 model, we use the mapping $f_u : 0 \rightarrow$ "premise", 1 → "claim" to associate the respective labels. In the T5 tokenizer, capitalized words may be split into two tokens. Thus, 0 and 1 are converted to "premise" and "claim" respectively.

[1] https://github.com/THUDM/ChatGLM-6B
[2] https://github.com/ProsusAI/finBERT
[3] https://huggingface.co/datasets/financial_phrasebank
[4] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html#sklearn.metrics.f1_score

Result	Prompt
TUA1-0 in AUI	Question: witch view does the following argument belong to? Premise or Claim? Argument:
TUA1-1 in AUI	Choose Premise or Claim:
TUA1-0 TUA1-1 in ARI	Below are two sentences that contain opinions. Please judge the logical relationship between the two sentences.

Table 2: Prompts for Submission Results.

Label	Meaning
0	There is **no detected** relation between the two sentences.
1	There is a **"Support"** relation from sentence 1 to sentence 2.
2	There is an **"Attack"** relation from sentence 1 to sentence 2.

Table 3: Meaning of each label in Argument Relation Identification.

2.2 Financial Phrasebank

The Financial phrasebank is a sentiment dataset consisting of sentences related to financial news. The dataset contains 4,840 English sentences, each classified as negative, neutral, or positive by five to eight annotators[5].

By first pre-training language models on either domain-specific data (Domain-Adaptive Pre-Training or DAPT) or data relevant to a specific task (Task-Adaptive Pre-Training or TAPT), enhanced performance on downstream tasks. Financial data, characterized by its strong professionalism and distinct data boundaries, is well-suited for such advanced pre-training.

Therefore, in addition to the original T5 model, In this paper fine-tunes the T5 model using the LoRA-based financial phrase bank dataset.

3 ARGUMENT RELATION IDENTIFICATION

The Argument Relation Identification subtask classifies the input text into three categories: Attack, Support, and No detected. An example data set is shown below:

> Sentence 1: So, just in terms of GDPR I kind of address that earlier, we do see that having an impact in Europe that we did see a reacceleration in growth in Europe, as we have lapped the initial implementation of GDPR.
>
> Sentence 2: So that 2019s promising, but we continue to see Europe growing just a bit slower than the rest of the regions
>
> Label : 2

The mapping of each label is shown in Table 3.

Similar to the Argument Unit Identification, experiments were conducted using the pre-trained models, FinBERT and T5. For T5, models were fine-tuned using the financial phrasebank dataset.

In this subtask, experiments were evaluated using the F1-Score. The highest F1-Score was achieved by the Fine-tuned T5 model. During the experiment with T5, labels 0, 1, and 2 were converted to "none", "support", and "attack" for label mapping. Using "no relation" or "unrelated" as mapping words would cause the tokenizer to split them into two tokens. By selecting "none" as the mapping, the outputs from the T5 model are always consistent with the mapping vocabulary. As a result, the final output label can be directly obtained without any additional decoding.

Additionally, this dataset is unbalanced. i.e., the number of each label included in the validation dataset is shown in Table 4.

Label	Sum	Rate
No Detected (0)	200	28.99%
Support (1)	482	69.86%
Attack (2)	8	**1.16%**

Table 4: The number of each label in the Argument Relation Identification validation data.

Table 4 reveals that the number of "Attack" labels is extremely small in the dataset. This is also the case for the training data. Machine learning with unbalanced datasets often results in decreased Precision and Recall, though experiments indicate that Accuracy is not significantly impacted. This explains why an imbalanced dataset adversely affects the evaluation in macro-F1, which incorporates both of these metrics in its calculations. Therefore, we have proposed two methods to address the problem of unbalanced datasets. The first method is the Weighted Random Sampler, and the second is the Cost-Sensitive Learning method. This chapter describes each method in detail and introduces FinBERT.

3.1 Weighted Random Sampler

This method was employed in experiments conducted with T5. In addressing unbalanced datasets, oversampling of minority labels can enhance prediction accuracy for those labels. During the division of training data into mini-batches, weights are applied to increase the probability of sampling minority labels, thereby mitigating the data imbalance within each mini-batch[6].

[5] https://huggingface.co/datasets/financial_phrasebank

[6] https://pytorch.org/docs/stable/data.html

3.2 Cost-Sensitive Learning

Experiments at FinBERT were conducted using the Cost-Sensitive Learning method. Cost-Sensitive Learning is a learning method that uses a specific cost in the machine learning process and takes that cost to improve classification accuracy. There are various methods for defining cost. In this experiment, we used a method that assigns a large cost to minority classes when they are misclassified. We tuned the training of the model by giving each class weights for the loss function called Cross-entropy Loss [5].

3.3 FinBERT

Finally, we introduce FinBERT. FinBERT is a machine learning model that takes BERT, a pre-trained model commonly used in natural language processing, and tailors it specifically for natural language processing tasks related to the financial sector. Two datasets were used for training, one being the Reuters TRC2 dataset and the other being the Financial phrasebank dataset for sentiment analysis. Financial documents often use unique vocabulary and expressions. By performing this fine tuning, FinBERT is able to understand expressions unique to the financial sector that cannot be successfully analyzed by previous natural language processing models [6].

4 EXPERIMENTS

In this section we describe the results we have submitted.

4.1 Submitted Results

In FinArg-1 of NTCIR-17, the TUA1 team submitted two results for the Argument Unit Identification subtask and three results for the Argument Relation Identification subtask. For each result, the methods used are listed in Table 5 [3].

Submission	Method	Micro -F1	Macro -F1	Weight -F1
TUA1-0 in AUI	T5	74.61	74.56	74.62
TUA1-1 in AUI	T5+ FP	76.37	76.36	76.38
TUA1-0 in ARI	T5	85.94	55.36	85.13
TUA1-1 in ARI	T5+ FP	85.65	61.50	84.86
TUA1-2 in ARI	Fin BERT +CSL	81.30	56.26	80.76

Table 5: Results of AUI and ARI submissions. FP stands for Financial phrasebank. CSL stands for Cost-Sensitive Learning.

Argument Unit Identification's TUA1-1 placed third among all submitted results. Also, TUA1-1 in Argument Relation Identification had the best results among all participants. For those results, the T5+ financial phrasebank method was employed. Thus, it was shown to be a superior method compared to other methods. Next, we describe the parameters used in each experiment.

4.2 Parameter

We present the parameters of the pre-trained model employed in our submission results. For the Argument Unit Identification subtask, the parameters can be found in Table 6, while those for the Argument Relation Identification subtask are detailed in Table 6.

Submission	Parameter			
	T5 size	Optimizer	Batch Size	lr
TUA1-0 in AUI	large	Adam	16	3e-4
TUA1-1 in AUI	large	Adam	16	1e-4
TUA1-0 in ARI	large	Adam	8	1e-4
TUA1-1 in ARI	large	Adam	8	1e-4
TUA1-2 in ARI	-	AdamW	64	1e-4

Table 6: Parameters for AUI and ARI submission results. lr stands for learning rate.

Various sizes of models exist in T5. They range from as small as 220 million parameters to as large as 11 billion. However, we conducted our experiments using the 770 million parameter model, termed Large. The primary reason for this choice was running costs. Models with larger parameters demand more GPU performance and take longer to run. Consequently, we experimented with both T5-Base and T5-Large. We chose to submit the results for T5-Large, as it achieved a better F1-score than T5-Base in our tests. We also experimented with and tuned other parameters under different conditions.

TUA1-2 is the submission result obtained using the FinBERT and Cost-Sensitive Learning methods. For this submission, the values of 1.8, 1.0, and 50.0 were assigned to weight-0, weight-1, and weight-2, respectively. These values indicate the costs associated with each label; a higher value implies a greater penalty added to the loss function upon misclassification. Initially, we established a cost of 1.0, drawing from the number of labels in each validation data set as presented in Table 4. We used the 'Support' label, which has the highest label count, as our based. Subsequently, we determined the cost for each label by taking the number of 'Support' labels as the numerator and the total count for each specific label as the denominator. After these calculations, we tuned the values, adopting the ones that yielded the best results for our final submission. Table 7 shows the changes in results due to differences in parameters.

Model	Weight-0	Weight-1	Weight-2	Macro -F1
FinBERT	-	-	-	0.518
FinBERT +CSL	1.5	1.0	41.3	0.560
FinBERT +CSL	1.5	1.0	62	0.573
FinBERT +CSL	1.8	1.0	50.0	**0.599**

Table 7: Parameters for Cost-Sensitive Learning. CSL stands for Cost-Sensitive Learning. Other parameters are the same as TUA1-2 in ARI.

4.3 Compare Experiments

All results submitted were experimented with either T5 or FinBERT, but we have experimented with the pre-trained method. The models used in the Argument Unit Identification subtask are listed in Table 8. The models used in the Argument Relation Identification subtask are also listed in Table 9.

Model	Accuracy
ChatGLM+p-tuning v2	0.764
T5	0.768
BERT	0.758
FinBERT	0.753
T5 + financial phrasebank	**0.774**

Table 8: Experimental Results for Argument Unit Identification.

We conducted experiments with various methods and discovered that T5, when fine-tuned using the financial phrasebank dataset, outperformed others in the Argument Unit Identification subtask.

Model	Macro-F1
T5	0.53
T5 + financial phrasebank	**0.58**
flan-T5	0.57
FinBERT + Cost-Sensitive Learning	**0.599**

Table 9: Experimental Results for Argument Relation Identification.

In the Argument Relation Identification subtask, T5, which was fine-tuned with the financial phrasebank datasets, and FinBERT, which used Cost-Sensitive Learning, both produced excellent results.

5 CONCLUSIONS

This paper presents and discusses the methods employed by the TUA1 team in both subtasks of FinArg-1. The experiments demonstrated that pre-trained models, based on Transformers, were fine-tuned using the dataset for both Argument Unit Identification and Argument Relation Identification. This fine-tuning revealed the potential for achieving high accuracy in text classification with Transformer-based models. In future research, we aim to explore the feasibility of achieving similarly high accuracy with large language models such as GPT-4 and LLaMa, which were not addressed in this study.

6 ACKNOWLEDGMENT

This research has been supported by the Project of Discretionary Budget of the Dean, Graduate School of Technology, Industrial and Social Sciences, Tokushima University.

REFERENCES

[1] Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li and Jun Zhao *FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining.* Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20), 2020

[2] Chung-Chi Chen, Hen-Hsen Huang and Hsin-Hsi Chen *From Opinion Mining to Financial Argument Mining.* Springer Nature 2021

[3] Chung-Chi Chen, Chin-Yi Lin, Chr-Jr Chiu, Hen-Hsen Huang, Alaa Alhamzeh, Yu-Lieh Huang, Hiroya Takamura, Hsin-Hsi Chen and Jun Zhao *Overview of the NTCIR-17 FinArg-1 Task: Fine-Grained Argument Understanding in Financial Analysis.* CIKM' 23, October 21–25, 2023

[4] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li and Peter J. Liu *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.* Journal of Machine Learning Research 21, 2020

[5] Elkan Charles. *The Foundations of Cost-Sensitive Learning.* Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI ' 01), 2001

[6] Dogu Tan Araci *FinBERT: Financial Sentiment Analysis with Pre-trained Language Modelsr.* arXiv:1908.10063v1 [cs.CL] 27 Aug 2019

[7] Fei Ding, Xin Kang, Yunong Wu, and Fuji Ren *TUA1 at the NTCIR-16 DialEval-2 Task.* Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies,2022

[8] Hanahira Shingo and XIN KANG *Resolving Context Contradictions in the Neural Dialogue System based on Sentiment Information.* The 7th International Conference on Machine Learning and Soft Computing (ICMLSC2023), 1-8, Jan. 2023

[9] Hu Junlin, XIN KANG, Shun Nishide and Fuji Ren *Text multi-label sentiment analysis based on Bi-LSTM.* Proceedings of 2019 6th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS2019), 16-20, Singapore, Dec. 2019

[10] XIN KANG, Shi Xuefeng, Yunong Wu and Fuji Ren *Active learning with complementary sampling for instructing class-biased multi-label text emotion classification.* IEEE Transactions on Affective Computing, Vol.14, No.1, 523-536, 2023

[11] Sun Yan, Fuji Ren, XIN KANG and Changqin Qua *Developing a Japanese Adverb-Emotion Corpus to Investigate the Effect of Adverbs in Japanese Sentence Emotion Classification.* International Journal of Advanced Intelligence (IJAI), Vol.8, No.1, 99-116, 2016