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/10.20736/0002001306

CYUT at the NTCIR-17 FinArg-1 Task2: A Quantitative Prompt Engineering Approach for Identifying Attack and Support Argumentative Relations in Social Media Discussion Threads

Shih-Hung Wu
Chaoyang University of Technology
Taiwan (R.O.C)
shwu@cyut.edu.tw

TSAI Tsung Hsun
Chaoyang University of Technology
Taiwan (R.O.C)
s11227607@gm.cyut.edu.tw

ABSTRACT

This paper reports our prompt engineering approach to the FinArg-1 task. In year 2023, we focus on task 2. Our system adopts the GPT3.5 generation model to evaluate the argumentative relations in social media discussion threads. We used three different prompts guide the GPT3.5 model to evaluate the degree of support or attack, we refer it as a quantitative approach. Our system then collected the score to make the final decision. The official results shows promising direction of using quantitative prompt engineering on argumentative relation identification.

KEYWORDS

Prompt Engineering, LLM, GPT3.5 model

TEAM NAME

CYUT

SUBTASKS

Task2: Identifying Attack and Support Argumentative Relations in Social Media Discussion Threads (Chinese)

1 INTRODUCTION

This paper reports our prompt engineering approach to the FinArg-1 task [1]. In year 2023, we focus on task 2. The FinArg-1 task 2 is a natural language processing (NLP) task that is part of the NTCIR (Natural Language Text Computing and Intelligent Retrieval) conference. The FinArg-1 task 2 is focused on identifying attack and support argumentative relations in social media discussion threads. Participants are asked to identify the argumentative relations (attack, support, or none) between two given social media posts. Each pair contains two posts, and labeled with support (1), attack (2), and none (0), the relations show that the second post support/attack the first post or no such relations. Details of the task description is in the task overview [1].

Finding the argumentative relation is particularly useful in the financial industry, where understanding financial arguments and their underlying reasoning is crucial for making informed investment decisions.

The task requires participants to develop algorithms and systems that can automatically extract financial arguments from text data. Our system adopts the GPT3.5 generation model to evaluate the argumentative relations in social media discussion threads. We used three different prompts guide the GPT3.5 model to evaluate the degree of support or attack, and our system collect the score to make the final decision. We call this prompts engineering approach with a quantitative prompt.

2 RELATED WORK

Argumentative relations identification is a field of study within computational linguistics that seeks to identify and analyze the structure of arguments within a text. This involves the detection and classification of claims, premises, and other argumentative components, as well as the relations between them. Large language models, such as OpenAI's GPT-3.5, can be used in this task. These models are trained on a vast amount of text data, allowing them to learn patterns in human language. They can identify argumentative structures in text based on the patterns they've learned.

Traditional ways to perform the task with large language models, including fine-tuning a pre-trained model: This involves taking a pre-trained language model, such as BERT or RoBERTa, and fine-tuning it on a labeled dataset of text data. The model is trained to predict the label for each piece of text, and the model's weights are adjusted based on the accuracy of its predictions. Fine-tune the model on a labeled dataset of text data, where each piece of text has been annotated with an argumentative relation label (e.g. attack, support, or none). The model then learns to predict the sentiment label for a given piece of text by analyzing the language and context in which the text is used.

Since 2020, prompt engineering is a popular method of designing effective prompts that can guide the GPT model to perform various natural language processing (NLP) tasks [2]. One such task is relation classification, which aims to identify the semantic relation between two texts from a predefined set of relations. By using prompt engineering, we can leverage the general knowledge and language skills of the GPT model to do the NLP task of relation classification without requiring any fine-tuning or additional data.

In-context learning is a technique that leverages the information available in the prompt to guide the generation of text by a deep learning model [3]. It can be useful for improving the quality, relevance and diversity of the generated text, as well as for reducing the amount of data and computation required for training the model.

In this article, we will introduce the concept of in-context learning, explain how it can be applied to deep learning prompt engineering, and provide some examples and best practices for using it in different domains and tasks.

Zero-Shot Prompting is a novel technique in deep learning prompt engineering that aims to generate natural language outputs without requiring any labeled data or fine-tuning [4][5]. It leverages the pre-trained language models and their ability to perform few-shot learning by providing them with carefully crafted prompts that encode the desired task and output format. Zero-Shot Prompting can be applied to various natural language processing tasks, such as text summarization, sentiment analysis, question answering, and more. In this article, we will introduce the basic principles of Zero-Shot Prompting, explain how to design effective prompts, and showcase some examples of Zero-Shot Prompting in action.

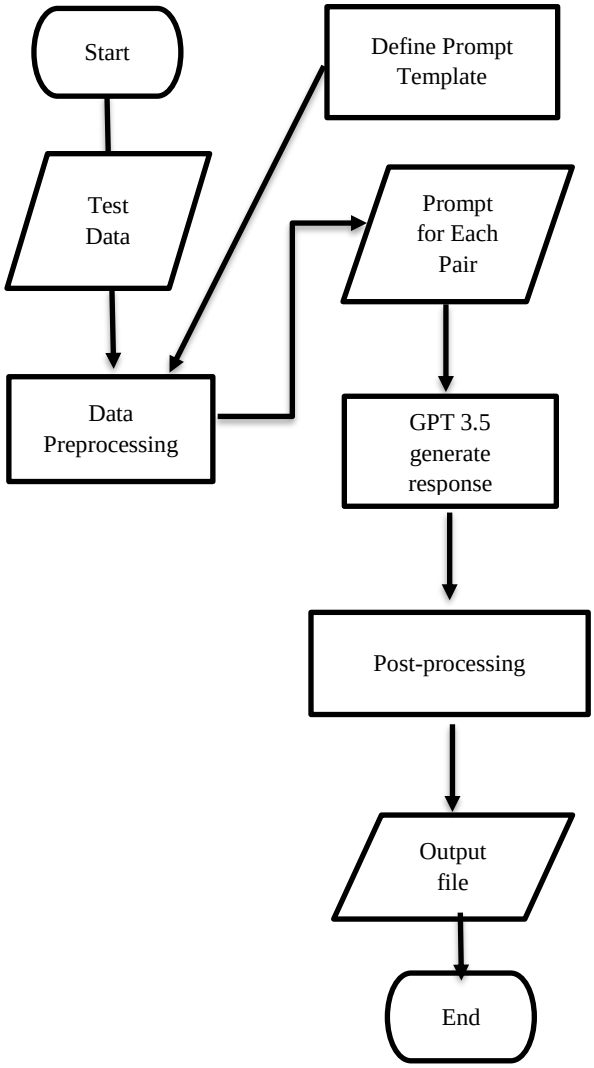

Figure 1: Flowchart of our Runs, each run with different prompt

3 Prompt Engineering Approach

This section describes our prompt engineering approach to the Task 2 problem. The general way of using prompt engineering for relation classification contains the following steps:

1. Define the task and the relations clearly and unambiguously.
2. Choose a suitable format and style for the prompt, such as natural language, cloze, or multiple choice.
3. Incorporate relevant keywords, cues, and examples that can help the GPT model understand the task and the relations.
4. Test and evaluate the prompt on different inputs and outputs, and refine it as needed.

We manually design our prompts 1 and 3 according to the zero-shot prompting and the in-context learning approach [6]. We hope that with the help of examples on none/attack/support the LLM might give better classification. On the other hand, we design the prompt 2 with our quantitative prompt engineering assumption. The assumption is that the system can analyze the response of LLM and find out the classification boundary by scoring each of the response. Figure 1 shows the flowchart of our approach. We defined three different prompt template as listed in Table 1. The data preprocessing module generates prompt for each input pair. The GPT 3.5 model receives the prompts and generates responses. The post-processing module collects all the response to generate output files for the sub-task.

3.1 The Prompts Used in Three Runs

The Prompt used in three runs are listed in Table 1. The run 1 and run 3 can be viewed as Zero-shot approach with a prefix prompts [7][8]. The prompts ask GPT model to generate classification results. Therefore the post-processing is straight forward, just collect the results generated by GPT. The run 2 named as quantitative approach is different. In run 2 the prompt asks the GPT model to generate a general response in text. In the post-processing module, a rule-based program is used to give the score on the degree of attack or support and then the system decide the prediction. The rule used in run2 is very rudimentary. The system count the number of "attack" or "support" to give scores and add 5 points or subtract 5 points from a default 50 point score.

Table 1: Prompt templates used in three runs model used in GPT3.5

run	Prompt template
1	1.編號 {ID}: 2.這是一個{sentence1}的帖子 3.編號 {ID}: 4.這是一個{sentence2}的帖子 5.這是無關的例子：[每張配現金 550 元加 17 股比去年多配 50 元這樣算好嗎?如果全配現金就好了高雄銀每張配現金六百不會稀釋股本但是小銀行是不是需要股本擴張 才能從事某些業務?,外資買超三個月了，今天衝高賣超 2000 多張，後續要觀察一下外資動作、如果回檔不多，或橫向移動，應該也不敢補了?

	6.這是支持的例子：[我們存股加碼隊加碼執行力很強、這篇新聞滿有趣的只提稅前沒提稅後大概是稅後不如預期所以故意不提。 7.這是攻擊的例子：[唉、快跌破 13 了正確說法已經比先前減資後重新上市價還低繼續跌破 12 就好笑了爛公司爛股票連財務操作也一樣爛慘字沒有下限。] 8.請根據這些資訊判斷帖子是支援、攻擊還是無關的： (1. Number {ID}: 2. This is a post of {sentence1} 3. Number {ID}: 4. This is a post of {sentence2} 5. This is an unrelated example: [Is it good to allocate 550 yuan in cash and 17 shares per ticket with 50 yuan more than last year?] If the cash is fully allocated, it is good, Kaohsiung Bank will not dilute the equity capital with 600 cash per sheet, but do small banks need equity expansion to engage in certain businesses?, foreign capital has been buying for more than three months, and today it is high to sell more than 2,000 pieces, and then it is necessary to observe the actions of foreign capital, if there is not much rollback, or lateral movement, should not dare to make up? 6. This is an example of support: [Our stock deposit plus team has strong execution, this news is full of interesting only mention before tax is not mentioned after tax, probably after tax is not as expected, so deliberately not mentioned.] 7. This is an example of an attack: [Alas, it is about to fall below 13, the correct statement is already lower than the re-listing price after the previous capital reduction, and it is funny to continue to fall below 12, rotten company, rotten stock, even financial operations, rotten miserable word, no lower limit.] 8. Based on this information, please judge whether the post is supportive, attacking, or irrelevant:)
2	這是一個{TestID}的帖子，請根據以下情境判斷是否支援、攻擊或無關： 帖子 1：{post1} 帖子 2：{post2} 情境 1：如果您認為帖子 2 支持帖子 1，請回復支持。 情境 2：如果您認為帖子 2 攻擊帖子 1，請回復攻擊。 情境 3：如果您認為帖子 2 與帖子 1 無關，請回復無關。 (This is a post of {TestID}. Please judge whether it is support, attack or irrelevant based on the following situations: Post 1: {post1} Post 2: {post2} Scenario 1: If you think Post 2 supports Post 1, please reply with support. Scenario 2: If you think Post 2 attacks Post 1, reply to the attack.

	Scenario 3: If you think Post 2 has nothing to do with Post 1, please reply Not related.)
3	1.編號 {ID}: 2.這是一個{sentence1}的帖子 3.編號 {ID}: 4.這是一個{sentence2}的帖子 5.這是無關的例子：[每張配現金 550 元加 17 股比去年多配 50 元這樣算好嗎?如果全配現金就好了高雄銀每張配現金六百不會稀釋股本但是小銀行是不是需要股本擴張 才能從事某些業務? 6.這是支持的例子：[我們存股加碼隊加碼執行力很強、這篇新聞滿有趣的只提稅前沒提稅後大概是稅後不如預期所以故意不提、對這死魚股來講，今天的確不尋常玩什麼把戲接著看下去、陽明海運今年居然衝到前 8 名，而且手持訂單比例居然比第一名的馬士基還多？、對這死魚股來講，今天的確不尋常玩什麼把戲接著看下去。 7.這是攻擊的例子：[唉、快跌破 13 了正確說法已經比先前減資後重新上市價還低繼續跌破 12 就好笑了爛公司爛股票連財務操作也一樣爛慘字沒有下限要有安全邊際價位時買入公司體質變壞才售出定存股每年還是得檢視中肛就是、請問陽明跌破圈購價 182，這樣還有意義嗎? 8.請根據這些資訊判斷帖子是支援、攻擊還是無關的： (1. Number {ID}: 2. This is a post of {sentence1} 3. Number {ID }: 4. This is a post of {sentence2} 5. This is an unrelated example: [Is it good to allocate 550 yuan in cash and 17 shares per ticket with 50 yuan more than last year?] It would be good if it was fully matched with cash, Kaohsiung Bank with 600 cash per sheet will not dilute the share capital, but do small banks need equity expansion to engage in certain businesses? 6. This is an example of support: [Our stock deposit plus code team plus code is very strong, this news is full of interesting only mention before tax did not mention after tax is probably not as expected after tax so deliberately not mentioned, for this dead fish stock, today is indeed unusual to play what trick to continue to read, Yangming shipping actually rushed to the top 8 this year, and the proportion of orders in hand is actually more than the first place Maersk? For this dead fish stock, today is indeed an unusual trick to continue watching. 7. This is an example of an attack: [Alas, it is about to fall below 13, the correct statement is already lower than the previous re-listing price after the capital reduction, continue to fall below 12, it is funny, rotten company, rotten stock, even financial operations, rotten miserable word, no lower limit, there is a margin of safety when the price buys the company's physique

	deteriorates before selling fixed deposit shares, every year still have to check the middle is, may I ask Yangming fell below the circle purchase price of 182, is this still meaningful? 8. Based on this information, please judge whether the post is supportive, attacking, or irrelevant：)

4 EXPERIMENT RESULTS

The evaluation metric for the task is the macro-average F1-score, which takes the precision and recall of three classes into account. The macro-average F1-score will give better score on the system with balanced performance on all three classes.

4.1 Official Data Distribution

The distribution of official data set is listed in Table 2. As we can see that the distribution of the data sets are similarly imbalanced among three datasets. In our approach, we use the dev set as our observation target during prompt design. We did not observe the training set and test set during system development. The distribution of data is a piece of information to find a set of result is plausible or not.

Table 2: The data distribution of training set, developing set, and test set

	# of 0 none	# of 1 support	# of 2 attack	total
dev	85	460	270	815
train	684	3676	2157	6517
test	85	460	270	815

	% of 0 none	% of 1 support	% of 2 attack
dev	10.44%	56.51%	33.05%
train	10.50%	56.41%	33.11%
test	10.44%	56.51%	33.05%

4.2 Official Results

Table 3 gives the official test result of our three runs. As we can see, the run 2 gives the best result among our three runs; we will give more analysis on the run 2 result. Table 4 shows the number of correct and wrong prediction of our run 2. The micro-f1 is 68.22%, not very far from the 2023 best micro-f1 71.66%. However, the marco-f1 is far lower than the best one. Table 4 shows the confusion matrix of run 2. In run 2 our system cannot find the none cases, the recall is 0%, therefore the macro-f1 is not good. There are some rooms to improve, so we designed an additional run in the following section.

Table 3: The official test results

Run	Micro-F1	Macro-F1	Weight-F1
CYUT-2	68.22%	49.62%	49.62%
CYUT-3	29.20%	23.45%	30.56%
CYUT-1	24.54%	20.94%	25.54%

Table 4: The confusion matrix of run 2

		System prediction		
		0(none)	**1(support)**	**2(attack)**
Official label	**0**	0	85	0
	1	0	372	87
	2	85	2	184

4.3 Examples of Run 2 Test Result

We give some examples on how our system label the sentence pairs. For ID 83, the correct result is none, but the score is 55, indicating support, so it is not correct. For ID 460: The correct result is support, the score is 45, indicating attack, so it is also not correct. For ID 681, the correct result is an attack, and the score is 45, indicating an attack, so it is a correct prediction.

Table 5: Examples in run 2 test results

ID	Sentences pair	System Score	Official label	Correct or not
83	"Post1": "今天發哥好強……看來是沒機會加碼了“ "Post2": "M大…請問反彈還有辦法彈回 430～435 嗎"	55	none	no
460	"Post1": "台中銀股價持續向 10 元靠近本益比 9 倍多可再持續買進“ "Post2": "想穩賺不賠就要等破 9 元，想買來生息像當定存, 10 元以下, 眼睛閉著買, 不要想太多! (以上,不負責任看法,當做我糊言亂語.)"	45	attack	no
681	"Post1": "一張不賣奇蹟自來“ "Post2": "一張不賣奇蹟自來 就是 當持股價值有 10 兆也不會賣的 因為奇蹟還沒來"	45	attack	yes

4.4 Additional Run

In run 2 our system cannot find the none cases, the recall is 0%. We find that the result is due to our simple rules of scoring. Therefore, we modified our prompt and conducted a new run. The major change is use the "反對(against)" install of "攻擊(attack)". The prompt used in the addition run is in Table 6, and the confusion matrix is in Table 7. In this run, our system can recognize 13 posts labeled 0 correctly, and the macro-average F1 improved to 50.10%.

Table 6: Prompt template used in additional run

這是一個{TestID}的帖子，請根據以下情境判斷是否支持、攻擊或無關：
貼文 1：{post1}
貼文 2：{post2}
情境 1：如果您認為貼文 2 支持貼文 1，請回覆支援。
情境 2：如果您認為貼文 2 反對貼文 1，請回覆攻擊。
情境 3：如果您認為貼文 2 與貼文 1 無關，請回覆無關。
你覺得 post2 中的語法或是單字對於 post1 是反對的感覺，還是支持的感覺，或是完全無關?

(This is a post from {TestID}. Please judge whether it is supportive, offensive or irrelevant based on the following scenarios:
Post 1: {post1}
Post 2: {post2}
Scenario 1: If you think Post 2 supports Post 1, please reply with support.
Scenario 2: If you think Post 2 is against Post 1, reply to the attack.
Scenario 3: If you think Post 2 has nothing to do with Post 1, please reply Not related.
Do you think the grammar or words in post 2 are against or supportive of post 1, or are they completely irrelevant?)

Table 7: The confusion matrix of additional run

		System prediction		
		0(none)	**1(support)**	**2(attack)**
Official label	**0**	**13**	72	0
	1	53	319	87
	2	97	2	172

5 CONCLUSIONS

We used three different prompts guide the GPT3.5 model to evaluate the degree of support or attack, and our system collect the score to make the final decision. The run 2 gives best result in official results. However, since the recall of class 0 is 0%, we found our weak point modified the prompt. An additional run is conduct and evaluated, the result shows that we fixed the zero recall problem and make improvement overall.

FinArg-1 task 2 is an important research direction in the field of NLP, and it has the potential to make a significant impact in the financial industry. By developing systems with LLM and prompt engineering, that can accurately extract financial arguments from text data, we can improve the efficiency and effectiveness of financial decision-making, and ultimately lead to better investment decisions in the future.

ACKNOWLEDGMENTS

This study was partially supported by the National Science and Technology Council under the grant number NSTC 112-2221-E-324-014.

REFERENCES

[1] 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

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[3] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, Z. Sui, A survey for in-context learning, arXiv preprint arXiv:2301.00234 (2022)

[4] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, Finetuned language models are zero-shot learners, arXiv preprint arXiv:2109.01652 (2021).

[5] Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In Proceedings of the Conference on EmpiricalMethods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19), Kentaro Inui, Jing Jiang, Vincent Ng, and XiaojunWan (Eds.). Association for Computational Linguistics, 3912–3921. https://doi.org/10.18653/v1/D19-1404

[6] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19). Association for Computational Linguistics, 2463–2473. https://doi.org/10.18653/v1/D19-1250

[7] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'21), Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 3045–3059. https://doi.org/10.18653/v1/2021.emnlp-main.243

[8] Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL/IJCNLP'21), Volume 1: Long Papers, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 4582–4597. https://doi.org/10.18653/v1/2021.acllong.353