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/10.20736/0002001312

SCUNLP-2 at the NTCIR-17 FinArg-1 Task: Enhancing Argumentative Relationship Recognition in the Classification Model with Language Generation Model Prompts

Han-Chiang Kao
Soochow University
Taiwan
woody850712@gmail.com

Hsin-Yun Hsu
Soochow University
Taiwan
crhhaa@gmail.com

Jheng-Long Wu
Soochow University
Taiwan
jlwu@gm.scu.edu.tw

ABSTRACT

While argument mining has significantly advanced across various domains, its application to financial discussions remains relatively unexplored. Our motivation for this research is rooted in the understanding that sentiment analysis alone may be inadequate when evaluating financial discussions, as the financial world is influenced by many factors intricately intertwined with the sentiments and opinions expressed by investors, analysts, and policymakers. To enhance the analysis of financial arguments, we incorporate GPT into the field of financial argument mining and design custom prompts. This unique integration allows us to generate labels and summaries for the arguments extracted from social media discussions. Our research results indicate that adding the generated labels in the regular mode achieved the highest validation set Marco-F1 score (66.39%). These findings contribute to a deeper understanding of argument mining in financial and social media discussions.

KEYWORDS

Argumentative Relations, Category classification, Text Generation, GPT, Social Media

TEAM NAME

SCUNLP-2

SUBTASKS

FinArg-1: Identifying Attack and Support Argumentative Relations in Social Media Discussion Threads (Social Media) (Chinese)

1 Introduction

Argument mining has witnessed significant advancements in various domains, and its application to financial discussions has yet to be explored. This research stems from recognizing that more than sentiment analysis is needed when evaluating financial discourse [1].

This paper presents our contribution to the NTCIR17-FinArg-1 shared task [3], where we focus on Fine-grained Argument Understanding in Financial Analysis. Our submission, titled "Identifying Attack and Support Argumentative Relations in Social Media Discussion Threads (Social Media)," represents an endeavor to bridge the gap in argument analysis within the financial context. To enhance the analysis of financial arguments, we introduce GPT (Generative Pre-trained Transformer) [2], an advanced language model, into the field of financial argument mining, and we design custom prompts to facilitate this integration.

Our primary goal is leveraging GPT to generate labels and summaries for arguments extracted from social media discussions, enabling a more nuanced and insightful understanding of investor sentiment and argumentative dynamics in the financial domain.

We will explore our research methods in Chapter 3, the role of GPT in enhancing argumentative relationship recognition, and the implications of our research for the broader field of financial analysis and sentiment assessment. Our findings will contribute to a deeper understanding of argument mining in financial and social media discussions, revealing the complex interplay between emotions, arguments, and financial decision-making.

2 Related Work

In recent years, there has been a growing convergence of natural language processing (NLP), argument mining, and financial analysis, which has emerged as a pivotal frontier in comprehending the intricate dynamics of financial markets and economic sentiment. The widespread use of social media platforms has enabled individuals worldwide to voice their opinions and perspectives concerning financial markets, economic policies, and investment choices. The exponential growth of user-generated content, primarily in the form of social media posts, has bestowed an excellent opportunity for researchers and analysts to delve into and gain insights into market dynamics in innovative ways. However, it's worth noting that while argument mining, the process of extracting structured information from unstructured text, has substantially advanced in various domains, its application to financial discussions remains an underexplored terrain [1]. Our motivation for conducting this research stems from our recognition that solely relying on sentiment analysis may prove inadequate when assessing financial discourse. The financial realm is influenced by many factors, many of which are intricately interwoven with the sentiments and viewpoints articulated by investors, analysts, and policymakers [1].

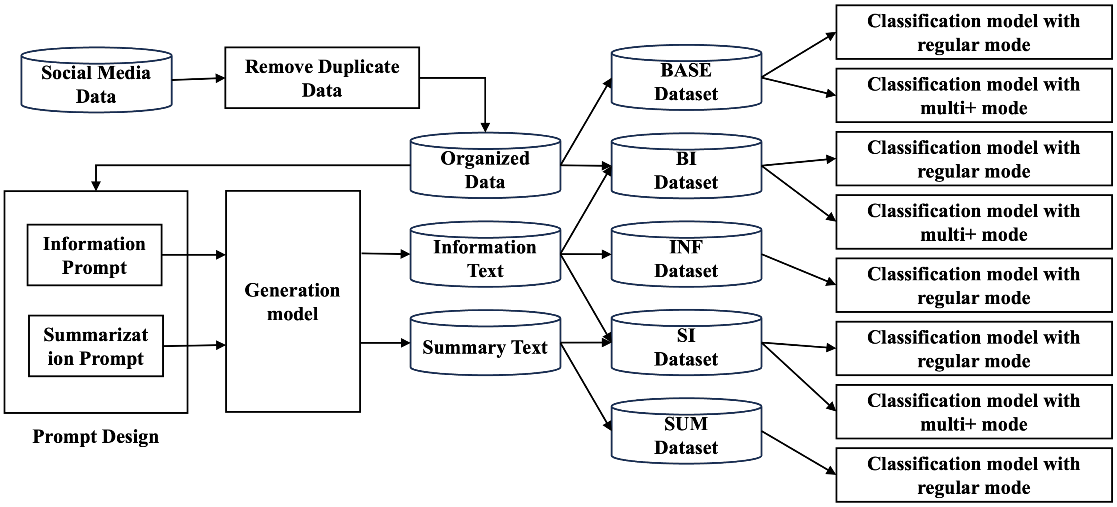

Social Media Data
Remove Duplicate Data
Organized Data
Information Prompt
Summarization Prompt
Prompt Design
Generation model
Information Text
Summary Text
BASE Dataset
BI Dataset
INF Dataset
SI Dataset
SUM Dataset
Classification model with regular mode
Classification model with multi+ mode
Classification model with regular mode
Classification model with multi+ mode
Classification model with regular mode
Classification model with regular mode
Classification model with multi+ mode
Classification model with regular mode

Figure 1: The flow chart of our study

Previous research compared content generated by ChatGPT with human-generated content and found that ChatGPT exhibited lower levels of bias and harmful information, almost no discriminatory dialogue, and a tendency to focus on the core issues [4]. The findings revealed that ChatGPT exhibited lower levels of bias and harmful information, almost no discriminatory dialogue, and a tendency to focus on the core issues. Additionally, in the realm of "emotion," ChatGPT often outperformed humans in expressing neutral emotions. Existing research in public sentiment analysis on stocks usually concentrates on the influence of human participants' emotions on the market. However, there is a noticeable gap in exploring the application and potential value of automatically generated language models like ChatGPT within this context. Our research seeks to gauge investor sentiment towards the stock market by analyzing social media posts [4]. We utilize the seven emotions labeled [5] (Happiness, Kindness, Sadness, Surprise, Fear, Disgust, and Anger) as a foundational framework for our question-and-answer template. We incorporate ChatGPT-generated emotional characteristics to determine whether this integration enhances the accuracy of sentiment analysis within the financial argument mining field.

As a result, we integrate GPT into financial argument mining and build custom prompts to enhance the recognition of financial argumentative relations. This approach represents a novel exploration, as existing research predominantly delves into the sentiment aspect of human participants' impact on the market, leaving a void in examining the potential value of automated language models. Consequently, our research bridges this gap by leveraging GPT and its generated emotional features to assess investor sentiment in social media discussions about the stock market. Our research objective is to amalgamate emotion labels and summaries generated by GPT with original social media posts to conduct an in-depth argument analysis.

3 Methods

Task 2 of FinArg-1 aims to identify argumentative relations within a financial social media dataset. This section will comprehensively explan how text generation models are leveraged to create new textual features to enhance the argumentative relations.

In the devised workflow, the initial step involves removing duplicate values from the original social media dataset, resulting in what we term the organized dataset. Next, we craft prompts designed for input into the language generation model. These prompts are enriched with financial context or commenter emotions, providing pertinent information for argumentative relations recognition. Subsequently, both the textual content of the organized dataset (the cleaned version of the original social media dataset) and the contents of the prompts are input into the language generation model. In our case, we employ the GPT text generation model to generate novel text features. These generated text features encompass both information text and summary text. These newly generated features are then integrated with the organized dataset, constructing five distinct datasets. Finally, these datasets are employed to train classification models, culminating in the ultimate classification model and conclusive outcomes for argumentative relations recognition. Additionally, when inputting these datasets into the classification models, we've designed a multi+ mode tailored for datasets exceeding 512 tokens. For a more detailed view, as presented in Fig 1.

3.1 FinArg-1 Social Media Dataset

In this classification task, we have a finance-related social media dataset consisting of three key fields: Post 1, Post 2, and Argumentative Relation. Post 2 responds to Post 1, and the Argumentative Relation indicates the stance taken by Post 2 toward Post 1, categorized as either support, attack, or none. Our objective is to use this dataset to recognize argumentative relations.

Word Count Observations. We observed that most BERT models have a token limit of 512 tokens. To address this limitation, we calculated the proportions and word counts of data exceeding 512 tokens in both the training, testing, and validation sets. You can find these statistics in Table 1, which includes the total counts for each dataset, the count of instances with more than 512 tokens, the maximum token count, and the proportion of data exceeding 512

Table 1: The total count of train/val/test datasets, the total count of more than 512 tokens, the maximum number of tokens, and the proportion of more than 512 tokens.

	Count	Count (greater than 512 tokens)	maximum word count	Proportion: Count of posts with greater than 512 tokens / Total
Training dataset	6518	821	2546	0.126
Testing dataset	815	124	1373	0.152
Validation dataset	815	108	1327	0.133

tokens. Notably, across all the train/val/test datasets, the proportion of data exceeding 512 tokens consistently exceeds 0.126. Consequently, the following experiments discuss how data that exceeds this token limit is handled, as outlined in sections 3.2 and 3.4.

Data Duplication. Two forms of data duplication are observed within the social media dataset: (1) Complete data duplication, where Post 1, Post 2, and Argumentative Relation are identical; (2) Duplication of Post 1 and Post 2, with differing Argumentative Relations. Different approaches were adopted to manage these distinct types of duplication.

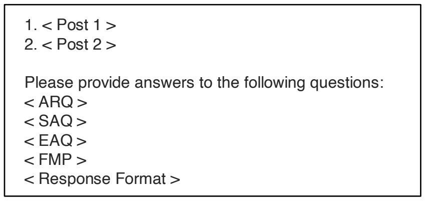
(1) Complete data duplication: We excluded five instances of wholly duplicated data. Duplicate data doesn't contribute to the improvement of classification models.

(2) Duplication of Post 1 and Post 2, with differing Argumentative Relations: In the training dataset, there were instances where posts in rows 1059 and 5572 were duplicated. Although the content of these posts was identical (Post1: "今天大漲百點發哥逆勢下跌..." and Post2: "IC 設計大廠聯發科今 7 日發布..."), they exhibited distinct argumentative relations (Row 1059: Attack, Row 5592: Support). While such cases were unexpected, after a discussion among three analysts, it was decided to label this scenario as an "attack."

3.2 Prompt Design

We devised novel prompts to guide the language generation model in providing information conducive to effectively distinguishing argumentative relations. This approach, which involves generating new features using prompts, applies to various language generation models, such as OpenAI's GPT and META's LLAMA. Our prompt content is categorized into two types: (1) Information Prompts: These prompts instruct the language generation model to generate insights about the commenter's perspective on the financial context and their emotional disposition. We integrated these inquiries into templates to elicit responses that could potentially aid the classification model. (2) Summary Prompts: In these prompts, the language generation model produced summaries of the posts. We ensured that the summaries remained within a 500-word limit derived from the 512-tokens limit of BERT. This word constraint was enforced to prevent the token count from exceeding the model's learning capacity, thus enabling us to perform subsequent classification tasks effectively. The purpose of using summaries is to assess whether focusing on the key content after summarization enhances the model's performance.

Information Prompts. The information Prompts comprise four distinct sections: Argumentative Relation Question, Sentiment Analysis Question, Emotional Analysis Question, and Financial Market Perspective. Additionally, we offer illustrative response demonstrations for the generated model. The placeholders <Post 1> and <Post 2> will be replaced with the content of the two original posts from the dataset. We will fill in different prompts for <ARQ>, <SAQ>, <EAQ>, <FMP>, and <Response Format>. As presented in Fig 2. Here's a detailed explanation of each section:

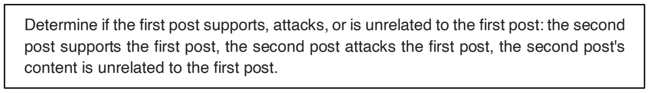


1. < Post 1 >
2. < Post 2 >

Please provide answers to the following questions:
< ARQ >
< SAQ >
< EAQ >
< FMP >
< Response Format >

Figure 2: Demonstrations of Information Prompt

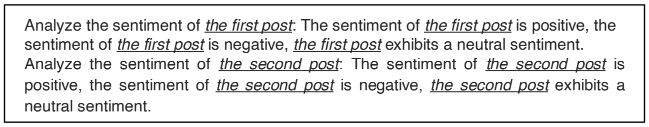
3.2.1 Section 1: Argumentative Relation Question. This section pertains to the recognition of the argumentative relation. This portion includes a single question that prompts GPT to determine the argumentative relations of the second post regarding the first post. This question aligns with the primary goal of the current classification task. Previous research has indicated that GPT's performance in stance detection surpasses traditional machine learning models [6]. In this case, we leverage GPT to provide us with the recognition of the argumentative relations after processing the posts. The template is presented in Fig 3.



Determine if the first post supports, attacks, or is unrelated to the first post: the second post supports the first post, the second post attacks the first post, the second post's content is unrelated to the first post.

Figure 3: < ARQ> in the Information Prompt.

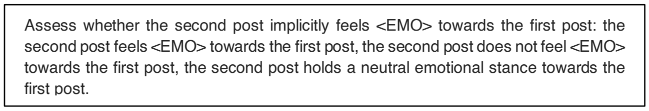
3.2.2 Section 2: Sentiment Analysis Question. This section revolves around conducting sentiment analysis on the two posts. This portion consists of two questions, each prompting GPT to determine the sentiment of each post individually—positive, negative, or neutral. These widely used sentiment questions offer the classification model more insights into the sentiments of the commenters. Previous research has indicated that the utilization of GPT yields superior performance in sentiment analysis compared to the original dataset [7]. The template is presented in Fig 4.



Analyze the sentiment of *the first post*: The sentiment of *the first post* is positive, the sentiment of *the first post* is negative, *the first post* exhibits a neutral sentiment. Analyze the sentiment of *the second post*: The sentiment of *the second post* is positive, the sentiment of *the second post* is negative, *the second post* exhibits a neutral sentiment.

Figure 4: < SAQ > in the Information Prompt.

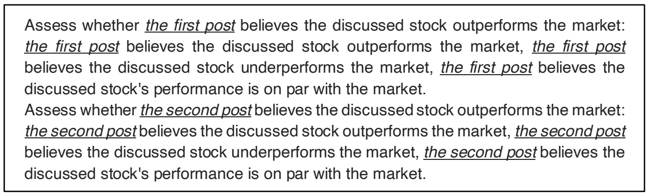
3.2.3 Section 3: Emotional Analysis Question. This section delves into the implicit emotional analysis of the second post in relation to the first post. It encompasses seven questions, each prompting GPT to provide an emotional analysis of how the second post implicitly expresses emotions towards the first post. These emotions are derived from previous research on emotions and the stock market [8] and are adapted to Chinese linguistic usage. The emotions include curiosity, joy, sadness, surprise, fear, disgust, and anger. The template is as follows, where <EMO> with each of the seven different emotions in each question. As presented in Fig 5.



Assess whether the second post implicitly feels <EMO> towards the first post: the second post feels <EMO> towards the first post, the second post does not feel <EMO> towards the first post, the second post holds a neutral emotional stance towards the first post.

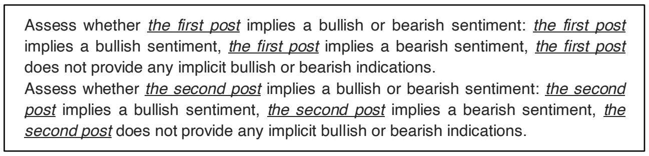
Figure 5: < EAQ> in the Information Prompt.

3.2.4 Section 4: Financial Market Perspective. This section delves into the sentiments expressed by the two posts towards the financial market. It comprises four questions, each prompting GPT to provide insights into how the two posts perceive the discussed stock compared to the broader market and whether they implicitly convey a bullish sentiment. We consider these views on the stock market to reflect the standpoints of the commenters, which aids in the recognition of the argumentative relations. The template is presented in Fig 6 and Fig 7.



Assess whether *the first post* believes the discussed stock outperforms the market: *the first post* believes the discussed stock outperforms the market, *the first post* believes the discussed stock underperforms the market, *the first post* believes the discussed stock's performance is on par with the market.
Assess whether *the second post* believes the discussed stock outperforms the market: *the second post* believes the discussed stock outperforms the market, *the second post* believes the discussed stock underperforms the market, *the second post* believes the discussed stock's performance is on par with the market.

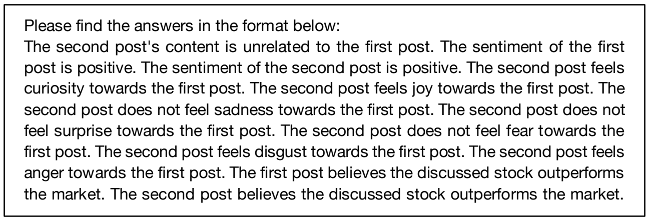
Figure 6: The first question of < FMP > in the Information Prompt.



Assess whether *the first post* implies a bullish or bearish sentiment: *the first post* implies a bullish sentiment, *the first post* implies a bearish sentiment, *the first post* does not provide any implicit bullish or bearish indications.
Assess whether *the second post* implies a bullish or bearish sentiment: *the second post* implies a bullish sentiment, *the second post* implies a bearish sentiment, *the second post* does not provide any implicit bullish or bearish indications.

Figure 7: The second question of < FMP > in the Information Prompt.

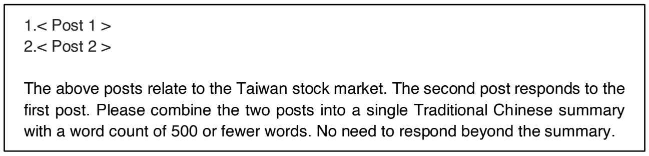
3.2.5 Response Format. Lastly, a response format for answering within the template. This approach reduces inconsistencies in template responses, providing a stable input for the classification model. The content is presented in the order of each preceding section, and we request GPT to respond in sequence to the provided questions. As presented in Fig 8.



Please find the answers in the format below:
The second post's content is unrelated to the first post. The sentiment of the first post is positive. The sentiment of the second post is positive. The second post feels curiosity towards the first post. The second post feels joy towards the first post. The second post does not feel sadness towards the first post. The second post does not feel surprise towards the first post. The second post does not feel fear towards the first post. The second post feels disgust towards the first post. The second post feels anger towards the first post. The first post believes the discussed stock outperforms the market. The second post believes the discussed stock outperforms the market.

Figure 8: < Response Format > in the Information Prompt.

The above four sections of questions, followed by the demonstrated template, constitute the complete content of the information prompt. At the beginning of the prompt, we provide GPT with the content of two posts. Each subsequent section builds upon the previous one, and the response format is added at the end. This entire content is input into the language generation model. The language generation model will respond with a comprehensive text that addresses each question and includes responses for all sections.
Summarization Prompts. Text summarization involves condensing two posts into a concise Chinese summary of 500 words or less. Aim to assess whether the summary generated by the language generation model enhances the classification model's ability to recognize argumentative relations. Previous research has indicated that, compared to the original text, GPT-generated summaries exhibit minimal differences perceptible to humans [9]. Since most text instances contain more than 512 tokens, they are constrained by BERT's token processing limitation. As a result, we hypothesize that replacing the original text with summaries will not have an adverse impact on classification performance. Furthermore, we ensure that the length of the summaries remains within 500 tokens, preventing the loss of essential information that may occur when exceeding the 512-token limit. The template is presented in Fig 9.



1.< Post 1 >
2.< Post 2 >

The above posts relate to the Taiwan stock market. The second post responds to the first post. Please combine the two posts into a single Traditional Chinese summary with a word count of 500 or fewer words. No need to respond beyond the summary.

Figure 9: Demonstrated Summarization Prompts.

3.3 Dataset Compilation

Section 3.2 introduced the designed prompts. Subsequently, we organize the processed original dataset and incorporate the prompts into the language generation model. This generates two sets of new text: information text and summary text. We reorganize and combine these texts with the organized dataset to form the following five datasets:

1. Base Dataset (BASE): This is the original FinArg-1-Social Media dataset without any generation model involvement. , with duplicate data removed.
2. Information Text Dataset (INF): Information text refers to the text generated by inputting the organized dataset into the language generation model using the Information Prompts. We compile these texts into a new dataset, excluding the organized dataset's text.
3. Summary Text Dataset (SUM): Summary text is generated by inputting the organized dataset into the language generation model using the Summarization Prompts. We compile these summaries into a new dataset, excluding the organized dataset's text.
4. Base + Information Text Dataset (BI): This dataset combines the information text with the organized dataset, creating a dataset with original text and text-generated information features. The aim is to provide the original posts with additional information beneficial for the classification model.

Table 2: Submissions for Three FinArg-1 Competitions.

Submission	Model	Dataset	Mode	Learning rate	Validation Marco-F1	Test Marco-F1
SCUNLP-2_SMDT_1	Hello-SimpleAI/chatgpt-detector-roberta-chinese	BASE	regular	1E-05	71.58%	59.61%
SCUNLP-2_SMDT_2	bert-base-chinese	BASE	regular	5E-06	71.53%	59.76%
SCUNLP-2_SMDT_3	Hello-SimpleAI/chatgpt-detector-roberta-chinese	BI	regular	5E-06	72.12%	66.39%

5. Summary Text + Information Text Dataset (SI): This dataset concatenates the generated summary text with the information text, creating a dataset with summaries and text-generated information features. This dataset explores whether classification can effectively be carried out using entirely generated text.

3.4 Handling Extensive Texts

Additionally, we've noticed that the Base, BI, and SI datasets in Section 3.3 still contain text passages that exceed the 512-token limit. This poses a challenge when inputting data into the classification model, as it cannot accommodate more than 512 tokens at once. To address this issue and ensure that the classification model doesn't miss out on valuable information beyond the 512-token threshold, we have designed a specialized input mode known as multi+ mode. We've categorized text input into two modes: regular and multi+.

In the regular mode, the original text is directly input into the classification model, with all text beyond 512 tokens being disregarded.

On the other hand, the "multi+" mode represents an enhanced input approach. In this mode, if the text length of a given row exceeds 512 tokens, we segment the text into new 512-token sections, starting from the 513th token and continuing this process throughout the entire text. If the original text and the segmented text do not collectively reach 512 tokens, we pad them with spaces (" ") to ensure each segment is a complete 512 tokens in length. After this padding, we have multiple segmented text sections, which we input into the classification model. The classification model produces vectors for each segmented section, and these vectors are summed to form the final classification. By adopting the multi+ mode, we aim to mitigate the loss of information in texts exceeding 512 tokens.

3.5 Pre-training Model

Following the design of these input modes, we utilized them to input the five datasets we curated in Section 3.3 into the classification models, resulting in the classification outcomes we sought.

We employed three distinct pre-training models for our experiments: bert-base-chinese, Hello-SimpleAI/chatgpt-detector-roberta-chinese (CDRC) [4], and papluca/xlm-roberta-base-language-detection (XRBLD) [10]. We chose “bert-base-chinese” as our initial baseline model due to its wide adoption as a Chinese language model. The selection of the other models was influenced by time constraints, primarily considering commonly used Chinese text classification models available on the Hugging Face website. Among these, CDRC and XRBLD emerged as candidates for our experimental models. CDRC incorporates text generated by GPT as a form of pre-training, which is advantageous for detecting our generated text. XRBLD demonstrates a notably high F1-Score during pre-training, specifically in the Chinese language context.

4 Experimental setup

This research employed the GPT-3.5 Turbo model developed by OpenAI to generate additional information, including information text and summary text, which were subsequently utilized to construct five distinct datasets. The experimental setup is divided into three aspects: model, dataset, and input mode. After configuring the aspects mentioned above and within time constraints, we conducted experiments using three different learning rate settings: 5e-6, 1e-6, and 5e-7. These experiments were carried out over 200 epochs. The model achieving the highest Marco-F1 score on the validation set among these trials was chosen as the representative model for this specific setup.

4.1 Evaluation Metrics

In this section, we elucidate how we evaluate the recognition of argumentative relations. Given that the ground truth of the test set in the FinArg-1 competition was not available before its conclusion, we employed the Macro-F1 score on the validation set to assess the accuracy of our trained classification model. Given the proportion of train/val/test datasets as 8/1/1, we consider the Macro-F1 score on the validation set to hold a certain reference value. All three results we provide are assessed using this metric as the evaluation criterion, as presented in Table 2.

Following the release of the FinArg-1 test dataset results, we utilized the test dataset's Marco-F1 score as an evaluation metric to gauge the models' capability in distinguishing argumentative relations.

5 Results

This research constructed five datasets and tested three classification models and two input modes on the existing dataset. The highest test Marco-F1 score when using the BI dataset with the regular mode of the CDRC model. As presented in Table 3.

Table 3: Experimental Results of Test Dataset Marco-F1 Scores.

	bert-base-chinese		CDRC		XRBLD	
	regular	multi+	regular	multi+	regular	multi+
BASE Dataset	0.609611	0.535087	0.617382	0.543349	0.632267	0.535006
INF Dataset	0.425922	-	0.451563	-	0.440516	-
SUM Dataset	0.467305	-	0.481383	-	0.474423	-
BI Dataset	0.671796	0.585185	**0.698785**	0.572415	0.674466	0.622763
SI Dataset	0.486866	0.502591	0.481728	0.495415	0.473756	0.491543

5.1 Results of Submission

In this section, we present a comparison of our submission results, as detailed in Table 2. The baseline setting, SMDT_2, utilizes the “bert-base-chinese” model in regular mode for the BASE dataset. Notably, SMDT_1 achieved the lowest test Marco-F1 at 59.61%, whereas SMDT_3 demonstrated the highest performance with a test Marco-F1 of 66.39%.

When contrasting SMDT_1 and SMDT_2, which employ different classification models and learning rate settings, we observe that SMDT_1's test Marco-F1 is slightly lower by 0.15% compared to SMDT_2. This discrepancy falls within an acceptable range and may be attributed to experimental variability, given that we conducted a single run for each configuration. It suggests that different classification models and learning rate settings had minimal impact on our experimental results.

In comparing SMDT_3 with both SMDT_1 and SMDT_2, SMDT_3 adopts distinct settings for the dataset and learning rate, resulting in a notable 6.78% higher test Marco-F1 compared to SMDT_1. Similarly, when contrasted with SMDT_2, which utilizes different classification models and dataset settings, SMDT_3 outperforms it by 6.63%. These findings underscore the improvement in classification model performance when augmenting the BASE dataset with additional information, highlighting the value of generated text in enhancing the model's capabilities.

5.2 Results on Different Datasets

In this section, we compare the performance of different datasets under the same mode. Within each classification model, under the regular mode, the highest Marco-F1 on the BI dataset is achieved (bert-base-chinese: 0.67, CDRC: 0.70, XRBLD: 0.67), while the lowest is observed on the INF dataset (bert-base-chinese: 0.43, CDRC: 0.45, XRBLD: 0.44). Under the multi+ mode, the highest Marco-F1 on the BI dataset is again achieved (bert-base-chinese: 0.59, CDRC: 0.57, XRBLD: 0.62), while the lowest performance is observed on the SI dataset (bert-base-chinese: 0.50, CDRC: 0.50, XRBLD: 0.49). We observe that the generated text consistently exhibits lower Marco-F1 scores when evaluated independently, whereas the BASE dataset offers more favorable information for the classification models.

Regarding the INF dataset, the generated text incorporates various additional information compared to the original text. Based on the results' Marco-F1 performance, the INF dataset shows lower scores than the BASE dataset. However, in the case of the BI dataset's Marco-F1 performance, it surpasses the BASE dataset. We speculate that the additional information introduced by the generated text may not have effectively enhanced the classification models' performance compared to the original text, leading to a performance decrease in the INF dataset. Nevertheless, combining the generated text with the original text in the BI dataset provides suitable cues for improved performance.

Regarding the SUM dataset, the generated text consists of summaries of the original text. Looking at the Marco-F1 performance, the SUM dataset shows lower scores than the BASE dataset. Moreover, the SI dataset's Marco-F1 performance is also lower than the BASE and BI datasets. Based on these results, we deduce that the summarization of the original text in the SUM dataset diminishes the crucial information required for the classification task. This suggests that generating summaries might not be the most effective approach in this experiment. Compared with the SI dataset, where the text summaries are concatenated with generated information text, there isn't a significant improvement in Marco-F1 performance compared to the SUM and INF datasets. This might indicate that combined generated text with summaries doesn't necessarily provide the classification model with more valuable information. Alternatively, it could signify that the generated text in both datasets offers similar information, preventing a substantial boost in Marco-F1 performance.

Regarding the SI dataset, the Marco-F1 performance is lower than the BASE and BI datasets. When comparing with the BI dataset, we suspect that the summarization text might have overlooked valuable information present in the original text that is advantageous for the classification task. This could be one of the reasons behind the relatively lower Marco-F1 score.

5.3 Results on Different Modes

In this section, we compare different input modes within the same dataset. Since the INF and SUM datasets did not use the multi+ mode, we won't discuss them in this section. In the BASE dataset, the regular mode shows higher Marco-F1 performance (bert-base-chinese: 0.61, CDRC: 0.62, XRBLD: 0.63). In the BI dataset, the regular mode exhibits higher Marco-F1 performance (bert-base-chinese: 0.67, CDRC: 0.70, XRBLD: 0.67). For the SI dataset, the multi+ mode achieves higher Marco-F1 performance (bert-base-chinese: 0.50, CDRC: 0.50, XRBLD: 0.49).

Regarding the BASE dataset, the regular mode outperforms the multi+ mode. The word count distribution of the BASE dataset, as presented in Fig 10(A), indicates a left-skewed distribution in both the 0-512 word range and the 513-1024 word range. When using the multi+ mode to perform the classification task on this dataset, there might be overuse of padding with spaces, leading to distortion in vector calculations. This could explain why the multi+ mode's performance is lower than the regular mode.

Regarding the BI dataset, the regular mode outperforms the multi+ mode. The word count distribution for the BI dataset is presented in Fig 10(B), where the majority of word counts fall within the range of 200-300 word in the 0-512 word interval, and in the 513-1024 word interval, the word count distribution also skews to the left. Similar to the BASE dataset, executing the classification task with the multi+ mode on this dataset may involve excessive padding with spaces, leading to lower performance than the regular mode.

Regarding the SI dataset, the multi+ mode slightly outperforms the regular mode. The word count distribution for the SI dataset is presented in Fig 10(C). In terms of word count statistics, the majority of data have word counts close to 512 words. This observation might be related to using word count restrictions during text generation, which effectively allows the multi+ mode to map text vectors without excessive padding.

Therefore, we believe that when utilizing the multi+ mode, it is essential to consider the dataset's specific characteristics. For example, view the SI Dataset, which comprises two texts generated by the language model, constrained by prompts and word limits to generate key content. In the regular mode, truncation of content may lead to the exclusion of crucial information and, in some cases, result in distorted vector calculations. Consequently, in this scenario, the regular mode's performance is inferior to that of the multi+ mode. This insight can serve as a valuable guide for achieving optimal performance with the multi+ mode by tailoring it to the dataset's unique attributes.

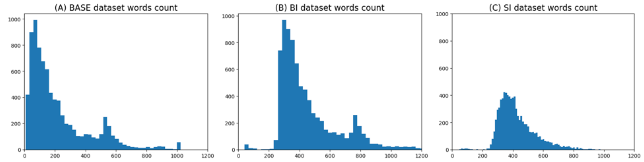


Figure 10: The word count distribution for the Base dataset, BI Dataset, and SI Dataset.

5.4 Results on Learning Rate

In this section, we discuss the varied Marco-F1 performances of the models at different learning rates. As presented in Fig 11, the test Marco-F1 across different learning rates for all experimental configurations. Each subplot within Fig 11 presents the experimental setup of the same dataset under the same input mode, showcasing different models' Marco-F1 results at varying learning rates.

Regarding the "bert-base-chinese" model, in the BASE dataset using the multi+ mode (Fig 11B), the BI dataset in regular mode (Fig 11E), and the SI dataset in multi+ mode (Fig 11H), larger learning rates correspond to higher Marco-F1 scores. Conversely, for the BASE dataset in regular mode (Fig 11A) and the SI dataset in regular mode (Fig 11G), larger learning rates are associated with lower Marco-F1 scores. For the INF dataset in regular mode (Fig 11C), the SUM dataset in regular mode (Fig 11D), and the BI dataset in multi+ mode (Fig 11F), the highest Marco-F1 scores are achieved when the learning rate is set to 1-e6.

Regarding the "CDRC" model, for the BI dataset in regular mode (Fig 11E), higher learning rates are associated with higher Marco-F1 scores. However, for the BASE dataset in regular mode (Fig 11A) and the BASE dataset in multi+ mode (Fig 11B), larger learning rates are linked to lower Marco-F1 scores. For the BI dataset in multi+ mode (Fig 11F) and the SI dataset in regular mode (Fig 11G), the highest Marco-F1 score is achieved when the learning rate is set to 1-e6. In the case of the INF dataset in regular mode (Fig 11C), the SUM dataset in regular mode (Fig 11D), and the SI dataset in multi+ mode (Fig 11H), the highest Marco-F1 score is attained with a learning rate of 5-e7, while the lowest score is observed at a learning rate of 1-e6.

Regarding the "XRBLD" model, for the INF dataset in regular mode (Fig 11C), the SUM dataset in regular mode (Fig 11D), the BI dataset in regular mode (Fig 11E), and the BI dataset in multi+ mode (Fig 11F), higher learning rates are correlated with higher Marco-F1 scores. However, for the BASE dataset in regular mode (Fig 11A), the BASE dataset in multi+ mode (Fig 11B), and the SI dataset in multi+ mode (Fig 11H), the highest Marco-F1 score is

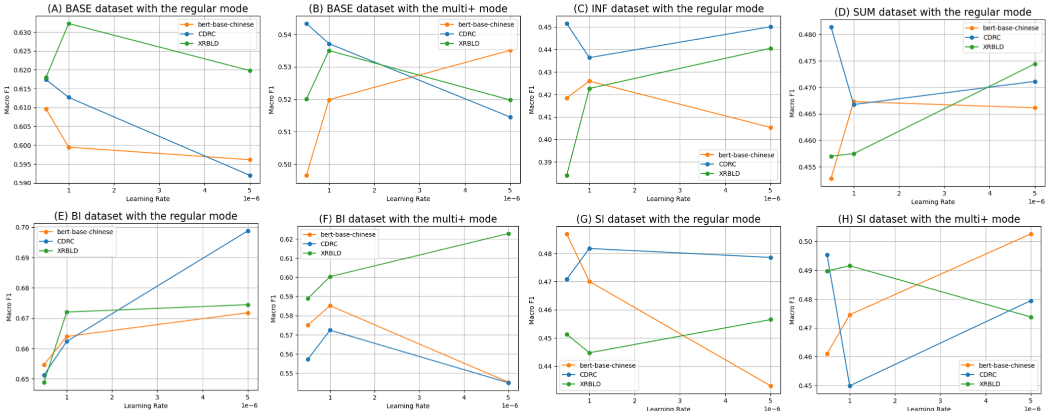


Figure 11: The test dataset's Marco-F1 across different learning rates for various experimental configurations.

achieved when the learning rate is set to 1-e6. For the SI dataset in regular mode (Fig 11G), the highest Marco-F1 score is achieved with a learning rate of 5-e6, while the lowest score is observed at a learning rate of 1-e6.

In most of our experimental configurations, we didn't observe a consistent trend in adjusting the learning rate concerning Marco-F1 across different models. However, it's worth noting that in the case of the BI dataset with the regular mode, we did notice that a higher learning rate led to a higher Marco-F1 score. The lack of a consistent trend in most settings could potentially be attributed to conducting single-run experiments for each configuration. In contrast, the BI dataset in regular mode consistently showed the highest performance among all Marco-F1 scores. Therefore, we conclude that this particular configuration achieved a level of stability in the classification model, resulting in a consistent trend with respect to the learning rate.

6 Conclusion

In this paper, we presented an approach to enhance the identification of argumentative relations using text generation techniques. We employed OpenAI's GPT API to generate two additional sets of text data apart from the original data: the Information Text dataset and the Summary Text dataset. These datasets were further constructed into five distinct collections: the original Base dataset (BASE), the Information Text dataset (INF), the Summary Text dataset (SUM), Base + Information Text Dataset (BI), Summary Text + Information Text Dataset (SI). Subsequently, we applied three classification models to perform the classification task and utilized two different input modes. The newly designed multi+ mode aimed to address the token limitations of classification models. We found that the highest validation set Marco-F1 score was achieved by the CDRC model on the regular mode of the BI dataset.

Regarding F1 scores regarding the datasets, the generated text datasets INF dataset and Sum dataset performed the worst, while the BI dataset outperformed the others. We believe that the generated text might have omitted certain information crucial for the classification task but could provide additional classification hints for the original text. Regarding input modes, the regular mode demonstrated better performance on the BASE and BI datasets than the multi+ mode, while on the SI dataset, the multi+ mode performed slightly better. We believe that the suitability of using the multi+ mode depends on the dataset's characteristics, suggesting that it might still offer better performance in specific scenarios. Regarding learning rates, we did not observe a consistent trend, which might be related to conducting a single run for each configuration. However, it's noteworthy that the BI dataset showed higher Marco-F1 with larger learning rates in regular mode. This result indicates this configuration's stability, suggesting that larger learning rates might benefit the classification model for this dataset.

In future work, we will continue to explore methods for leveraging generated text to enhance the performance of classification tasks.

LIMITATIONS

We acknowledge that the primary limitations of our approach stem from constraints related to time and budget. In terms of language generation models, understanding whether the generated text is beneficial for classification purposes requires additional resources for experimentation on platforms like GPT. However, due to budget constraints, we did not explore other language generation models, such as Google's Bard or Meta AI's LLaMA, potentially limiting the generalization of our results to GPT-generated text. Regarding the choice of classification models, our selection was constrained by time, leading us to consider only three representative models. Further research should investigate the performance of other models in this context. In the case of generated templates, we aimed to assess the utility of generated text for classification tasks. However, due to time and budget limitations, we performed single iterations of text generation for each template and did not deeply explore the quality of the generated text. Future research should delve into more extensive template designs for financial topics to better understand the impact of generated text quality on classification tasks.

REFERENCES

[1] Alexander Hogenboom, Frederik Hogenboom, Uzay Kaymak, Paul Wouters and Franciska de Jong, 2010. Mining economic sentiment using argumentation structures. In: *Trujillo, J., et al. Advances in Conceptual Modeling – Applications and Challenges. ER 2010. Lecture Notes in Computer Science*, vol 6413. Springer, Berlin, Heidelberg. DOI: https://doi.org/10.1007/978-3-642-16385-2_25

[2] Alec Radford, Karthik Narasimhan, Tim Salimans and Ilya Sutskever, 2018. Improving language understanding by generative pre-training. In *OpenAI Blog*.

[3] Chung-Chi Chen, Chin-Yi Lin, Chr-Jr Chiu, Hen-Hsen Huang, Alhamzeh, Alaa, Yu-Lieh Huang, Takamura, Hiroya and Hsin-His 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*, Tokyo, Japan.

[4] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue and Yupeng Wu, 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. DOI: https://doi.org/10.48550/arXiv.2301.07597

[5] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/P17-1152

[6] Bowen Zhang, Daijun Ding and Liwen Jing, 2022. How would stance detection techniques evolve after the launch of chatgpt? DOI: https://doi.org/10.48550/arXiv.2212.14548

[7] Kiana Kheiri and Hamid Karimi, 2023. Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning. DOI: https://doi.org/10.48550/arXiv.2307.10234

[8] Wenhao Chen, Yi Cai, Kinkeung Lai and Haorane Xie, 2016. A topic-based sentiment analysis model to predict stock market price movement using Weibo mood. In *Web Intelligence* (Vol. 14, No. 4, pp. 287-300). IOS Press. DOI: https://doi.org/10.3233/WEB-160345

[9] Tanya Goyal, Junyi Jessy Li and Greg Durrett, 2022. News summarization and evaluation in the era of gpt-3.DOI: https://doi.org/10.48550/arXiv.2209.12356

[10] 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. DOI: https://doi.org/10.48550/arXiv.1911.02116