

IMNTPU at the NTCIR-18 FinArg-2: Fine-Tuning and Prompt-Based Learning for Temporal Argument Detection and Claim Validity Assessment

Bor-Jen Chen
Information Management
National Taipei University
New Taipei City, Taiwan
s711336103@gm.ntpu.edu.tw

Wen-Hsin Hsiao
Information Management
National Taipei University
New Taipei City, Taiwan
s711336107@gm.ntpu.edu.tw

Jun-Yu Wu
National Taipei University
New Taipei City, Taiwan
s411084021@gm.ntpu.edu.tw

Cheng-Yun Wu
Information Management
National Taipei University
New Taipei City, Taiwan
s711336101@gm.ntpu.edu.tw

Min-Yuh Day*
Information Management
National Taipei University
New Taipei City, Taiwan
myday@gm.ntpu.edu.tw

ABSTRACT

The increasing availability of financial texts from earnings conference calls (ECCs) and social media has created a need for advanced natural language processing (NLP) techniques to extract meaningful insights. This study develops a classification framework that integrates fine-tuning and prompt-based learning to improve financial argument classification. We apply this framework to two tasks from the NTCIR-18 FinArg-2 competition: detecting temporal references in ECCs and assessing the validity period of claims in social media. Encoder-based models are fine-tuned for structured classification, while decoder-based models leverage both fine-tuning and prompt-based learning. Data augmentation techniques enhance model generalization, and performance is evaluated using Micro-F1 and Macro-F1 scores. The primary contribution of this research is demonstrating how fine-tuning and prompt-based learning can complement each other in financial NLP. By optimizing classification strategies, this study provides insights for improving argument analysis in financial applications, benefiting researchers, practitioners, and FinTech developers.

KEYWORDS

Financial Argument, Fine-tuning, Prompt-based Learning, Earnings Conference Calls (ECCs), Social Media.

TEAM NAME

Graduate Institute of Information Management NTPU NTCIR team

SUBTASKS

Detection of Argument Temporal References (English)
Assessment of the Claim's Validity Period (Chinese)

1 INTRODUCTION

Recent advancements in financial technology (FinTech) and the increasing availability of textual data from financial reports, earnings conference calls, and social media platforms have driven the need for improved natural language processing (NLP) techniques in financial analysis. Financial argumentation, particularly in earnings conference calls (ECCs) and investor discussions, provides critical insights into market trends and corporate performance. Extracting temporal references from financial texts is essential for analyzing historical trends and assessing their relevance to decision-making. Similarly, determining the validity period of investment claims in social media discussions is crucial for evaluating how long an opinion or forecast remains applicable. Both tasks present challenges due to the implicit nature of temporal cues and the dynamic nature of financial discourse.

With the rise of transformer-based language models, NLP research has evolved from traditional text classification techniques to more sophisticated fine-tuning and prompt-based learning approaches. Encoder-based models, such as BERT-based architectures, typically rely on fine-tuning with task-specific labels to improve performance on structured classification tasks. In contrast, decoder-based models, like GPT-4o Mini, can leverage fine-tuning alongside prompt-based learning, allowing them to generate structured responses based on in-context learning. This study explores how these two learning paradigms can be effectively applied to financial argument classification.

Given these technological advancements, this research aims to investigate the effectiveness of fine-tuned encoder-based models compared to fine-tuned decoder-based models using different prompting strategies. Specifically, this study focuses on two tasks introduced in the NTCIR-18 FinArg-2 competition: Detection of Argument Temporal References in Earnings Conference Calls (ECCs) and Assessment of the Claim's Validity Period in Social Media. The first task requires classifying temporal references in financial statements based on claims and premises, while the second task predicts the validity duration of financial claims in investor discussions. Through a comparative analysis of fine-

tuning and prompt-based learning, this study evaluates the role of data augmentation, model adaptation, and contextual learning in financial text classification.

In the subsequent portions of this document, Chapter 2 reviews related work 2 provided an overview of the related work to this study. Chapter 3 describes the methodology, including data preprocessing, model fine-tuning, and prompting techniques. Chapter 4 presents the experimental results, and Chapter 5 summarizes the findings and discusses future directions.

2 RELATED WORK

2.1 Temporal Inference of Financial Arguments

Temporal inference in financial texts is essential for understanding how time-referenced arguments influence market perceptions. In Earnings Conference Calls (ECCs), argument quality is assessed based on specificity, persuasiveness, and temporal references, where past financial data enhances credibility over forward-looking statements [1]. In social media investment discussions, impact duration estimation has been explored to classify whether financial events have short-term or long-term effects on stock movement. A pre-finetuning approach using investor-labeled duration data has been shown to improve stock prediction accuracy by incorporating temporal awareness [2]. These studies highlight the challenge of accurately modeling time-sensitive financial language in NLP applications.

2.2 Large language model (LLM)

Large Language Models (LLMs) represent a transformative advancement in natural language processing (NLP), significantly enhancing AI’s ability to understand and generate text. Encoder-based models, such as BERT, leverage bidirectional contextual embeddings to improve tasks like text classification and named entity recognition, though fine-tuning requires substantial labeled data and computational resources [3, 4]. Optimized variants like RoBERTa extend training durations and, when integrated with LSTM, enhance sentiment analysis performance [5].

In contrast, decoder-based models like GPT-4 employ an autoregressive architecture designed for text generation. With reinforcement learning from human feedback (RLHF), GPT-4 improves factual accuracy and response alignment, making it particularly effective for classification tasks with minimal labeled data [6]. Structuring input prompts has been shown to enhance multi-label classification accuracy and model consistency across complex NLP tasks [7].

2.3 Fine-tuning and Prompt Engineering

Recent advancements in transformer models have driven the adoption of fine-tuning and prompt-based learning for NLP tasks. Fine-tuning is a key approach for encoder-based models like BERT, where task-specific labeled data is used to update all model parameters, improving classification performance by capturing domain-specific patterns [8]. Optimizing hyperparameters and applying strategies like domain-specific pre-training further enhance generalization [9].

Decoder-based models like GPT leverage both fine-tuning and prompt-based learning. Prompt engineering enables classification with minimal labeled data using structured input templates [10].

Few-shot learning with well-designed prompts achieves performance comparable to fine-tuned non-generative masked language models (MLMs), demonstrating the effectiveness of in-context learning as an alternative to traditional model adaptation[11]. Additionally, systematic prompt tuning and structured prompt patterns further refine classification accuracy[12].

Overall, optimizing large language models involves a combination of fine-tuning and prompt engineering, each offering distinct advantages. Fine-tuning enhances structured classification stability, while prompt-based learning improves adaptability with fewer labeled samples. Integrating both approaches enhances NLP model efficiency and flexibility [13].

2.4 Data Augmentation

Data augmentation enhances NLP model performance, particularly in low-resource scenarios. Traditional methods like synonym replacement, word swapping, and deletion increase data diversity but often introduce noise [14]. Transformer-based approaches, such as BERT and GPT, offer more context-aware augmentation, improving classification accuracy, especially in long-text datasets [15].

Paraphrasing techniques using GPT-2 and CVAE generate diverse training samples while preserving label consistency, benefiting intent classification and slot filling [16]. In sentiment and financial text classification, GPT-generated synthetic data mitigates class imbalance, significantly improving model performance for underrepresented categories [17, 18].

Prompt-based augmentation refines data quality by generating semantically precise samples. GPT-3, through structured prompts, enhances event relation classification and improves low-resource categories more effectively than traditional methods [19]. A systematic review highlights GPT-based augmentation’s advantages in generating diverse, contextually relevant text while stressing the need for data filtering [20].

3 METHODS

3.1 Proposed Research Architecture

This study proposes an architecture that integrates fine-tuning and prompt-based learning to enhance transformer models’ performance in the NTCIR-18 FinArg-2 tasks. The workflow consists of data preprocessing, training strategies, model selection, and evaluation metrics, as illustrated in Figure 1.

The process begins with data preprocessing, where the FinArg-dataset undergoes cleaning and standardization to ensure consistency. Additionally, data augmentation is applied to both encoder-based and decoder-based models to improve model generalization, particularly in the Earnings Conference Call (ECC) subtask.

In the training stage, encoder-based models (e.g., BERT, RoBERTa, DistilBERT) undergo fine-tuning with task-specific labels to capture financial argumentation patterns. Decoder-based models (e.g., GPT-4o Mini) use both fine-tuning and prompt-based learning, where fine-tuning adapts the model to the task, and prompt-based learning enables structured predictions via in-context learning.

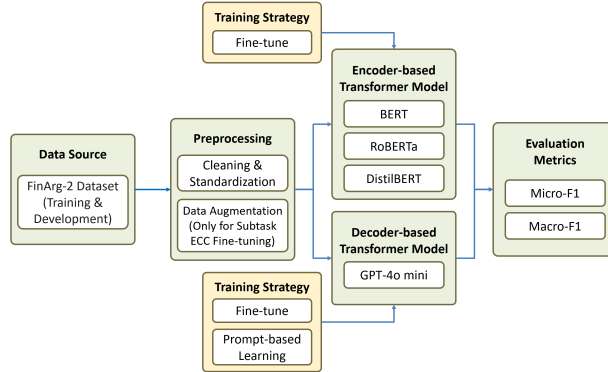


Figure 1: Proposed Research Architecture

For model selection, encoder-based models are chosen for their strength in text classification and feature extraction, while decoder-based models offer greater flexibility in handling various input formats.

Finally, models are evaluated using Micro-F1 and Macro-F1 scores, which align with the official ranking criteria of the FinArg-2 competition. The following sections provide a detailed breakdown of the implementation for each subtask.

3.2 Detection of Argument Temporal References (Earnings Conference Call)

The goal of this subtask is to classify argument claims based on their temporal reference using associated premises. Each claim is assigned one of three labels: 0 (No time reference), 1 (Long past, more than half a year), and 2 (Short past, within half a year, including the current or previous two quarters). Given the nature of financial earnings discussions, correctly identifying temporal references is crucial for analyzing historical trends and their relevance to decision-making. The challenge lies in distinguishing implicit and explicit temporal cues within the argument and supporting premises.

3.2.1 Data Description and Augmentation. The dataset provided by the competition organizers consists of training and validation sets, each containing labeled claim-premise pairs with associated temporal references. Each sample includes a `claim_text`, which represents the argument being made, and `premise_text`, which provides supporting context. Additionally, the dataset includes the year and quarter of the financial discussion and a label indicating the temporal classification.

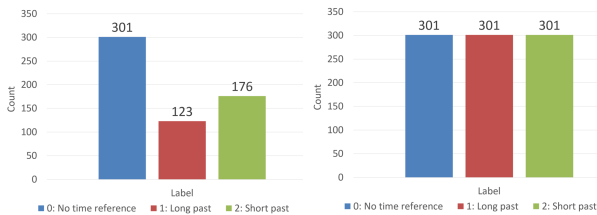


Figure 2: Label distribution diagram before and after data augmentation

The training dataset originally contained 600 samples, with an imbalanced distribution across the three labels, which could lead to model bias. To address this issue, data augmentation was applied to balance the label distribution, ensuring equal representation across all labels. After augmentation, the training set contained 903 samples, with each label having an equal number of instances. The validation dataset consists of 150 samples and was kept in its original form without augmentation, ensuring that model evaluation reflects real-world data distributions.

To generate additional samples, GPT-4o mini was used to create semantically equivalent variations of the original premises while preserving their meaning. The augmentation process involved extracting premises from the dataset and rephrasing them using a controlled generation prompt. The following prompt was employed to guide GPT-4o mini in generating alternative expressions:

"The task is to generate one alternative expression for the following premise while preserving its original meaning, tone, and accuracy.

Requirements:

1. Rewrite the premise into one alternative expression.
2. Retain critical terms and other financial-specific terminology.
3. Ensure the meaning, facts, and intent remain unchanged.
4. Avoid introducing new facts or speculative content.
5. Keep the tone formal and concise."

By using this structured prompt, the generated sentences maintain consistency with the original data while introducing slight variations in phrasing. To ensure data quality, a manual sampling review of 10% of the augmented samples was conducted, where randomly selected entries were checked for semantic consistency and factual accuracy. The augmented dataset provides a more balanced representation of each class, allowing the model to generalize better across different temporal references.

3.2.2 Fine-tuning Encoder-based Transformer Model

3.2.2.1 Input Configurations. To determine the most effective input representation, we experimented with three different configurations: using all available fields (`claim_text`, `premise_text`, `year`, `quarter`), using only `claim_text` and `premise_text`, and using only `premise_text`. Each configuration was tested to assess its impact on classification performance.

Based on initial observations, using only `claim_text` and `premise_text` provided better performance, while adding `year` and `quarter` did not significantly improve classification accuracy. Therefore, subsequent fine-tuning experiments adopted the `claim_text + premise_text` configuration.

3.2.2.2 Fine-tuning Procedure. This study fine-tuned six transformer-based models: BERT, RoBERTa-base, RoBERTa-large, DistilBERT, ALBERT, and FinBERT, using pretrained weights and adapting them for the argument temporal classification task. All models used their corresponding tokenizers to preprocess textual data. Tokenization included adding special tokens ([CLS], [SEP]), truncating sequences to a maximum length of either 128 or 256 tokens and applying padding to maintain uniform input dimensions. The tokenized sequences were then converted into tensor format, including input IDs, attention masks, and class labels, for model training.

The training process was optimized using the AdamW optimizer with cross-entropy loss. Each model was fine-tuned for 3 to 6 epochs, with early stopping based on validation set performance to prevent overfitting. Micro-F1 and Macro-F1 scores were monitored during training to ensure class balance. To further improve generalization, gradient clipping and weight decay (0.01) were applied, along with dropout layers in the transformer models.

A hyperparameter search was conducted to determine optimal training configurations. Table 1 summarizes the range of hyperparameters explored:

Table 1: Fine-tuning Encoder-based Transformer Model Hyperparameter Settings

| Hyperparameter | Value |
|----------------|--------------------------|
| Learning rate | 1e-5, 1.5e-5, 3e-5, 5e-5 |
| Max Length | 128, 256 |
| Batch Size | 16, 32 |
| Epochs | 3, 4, 5, 6 |

3.2.2.3 Model Selection. Based on initial observations, RoBERTa-base and DistilBERT demonstrated stronger performance compared to other models and were therefore selected as the final models for submission. A more detailed analysis of the results is provided in Section 4.1

3.2.3 Fine-tuning GPT-4o-mini and Prompt-based Learning

3.2.3.1 Fine-tuning GPT-4o-mini. To enhance the model's classification performance, GPT-4o-mini was fine-tuned using OpenAI's official platform. The fine-tuning dataset consisted of 900 augmented samples, formatted in JSON according to OpenAI's specifications.

The fine-tuning process was conducted with a learning rate of 0.1, a batch size of 4, and for 3 epochs.

Table 2: Fine-tuning GPT-4o-mini Hyperparameter Settings

| Hyperparameter | Value |
|----------------|-------|
| Learning rate | 0.1 |
| Batch Size | 4 |
| Epochs | 3 |

3.2.3.2 Prompt-based Learning. To evaluate the effectiveness of fine-tuning, both the pretrained and fine-tuned GPT-4o-mini models were tested under different prompting strategies: zero-shot, one-shot, three-shot, and six-shot learning.

Zero-shot learning provided only an instructional prompt:

Task: Classify the given text into one of the following categories:

0: No time reference (the text does not contain any explicit time indication)

1: Long past (the text refers to an event that occurred more than six months ago)

2: Short past (the text refers to an event that occurred within this quarter or the past two quarters)

Now, classify the following text:

Claim:

Premises:

Label:

In one-shot, three-shot, and six-shot learning, the prompt was extended with labeled examples before presenting the test instance. For example, a three-shot prompt included: You will first see three examples demonstrating how to classify text. Then, you will be given a new text to classify.

This was followed by labeled examples:

You will first see three examples demonstrating how to classify text.

Then, you will be given a new text to classify.

Example 1:

Claim: "It's an increasingly important use case for us."

Premises:

- "And people are already using Facebook to share during real-time events."

- "So this gives people to share, a place to share that one event and participate in it."

*Label: **0***

Example 2:

Claim: "So we're happy with the engagement that customers have."

Premises:

- "Well, ultimately, I'll step back and say one of the main things we look out on Prime Video is customer usage patterns and in 2016 we had a doubling of Prime hours for video, music and reading."

*Label: **1***

Example 3:

.....

By comparing the classification results of fine-tuned and non-fine-tuned models across different prompting strategies, this study examines the impact of fine-tuning and contextual information on performance. The results are presented in Section 4.1.

3.3 Assessment of the Claim's Validity Period (Social Media)

The objective of this Social Media subtask is to assess the validity period of a given claim. Participants are required to predict the validity duration for which a claim remains valid based on the provided dataset.

3.3.1 Data Description. The dataset provided by the FinArg 2 organizer[21] consists of both a training set and a development set, each comprising multiple JSON records. Each record contains two fields: "text" and "Label_Duration." The "text" field represents investors' opinions, presented in Chinese, while "Label_Duration"

is classified into three categories: "Longer than 1 week", "Within 1 week", and "Unsure".

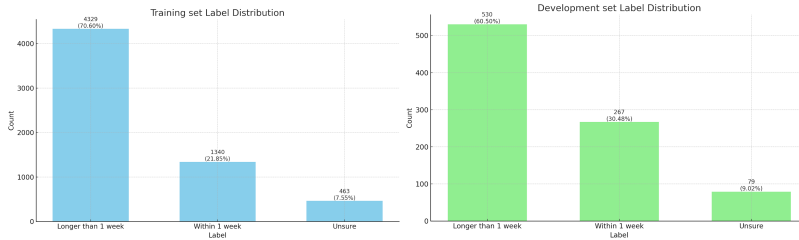


Figure 3: Label distribution diagram

The original training set consists of 6,132 records, with the following distribution of labels: 4,329 records labeled as "Longer than 1 week", 1,340 as "Within 1 week", and 463 as "Unsure". For the development set, a total of 876 records are included, with the following label distribution: 530 records classified as "Longer than 1 week", 267 as "Within 1 week", and 79 as "Unsure". The label distribution of the training and validation sets is visually presented in Figure 3.

In summary, the dataset for this subtask contains a substantial number of labeled records; therefore, we attempt to fine-tune using the original training dataset.

3.3.2 Encoder-based Transformer Model Fine-tuning

3.3.2.1 Label Encoding. In this subtask, Label Encoding is utilized to convert textual duration labels into numerical values, ensuring compatibility with the model's input format. This transformation facilitates efficient processing and enhances the representation of categorical information within the model.

The textual duration labels are encoded into numerical values, where "Within 1 week" is mapped to 0, "Longer than 1 week" to 1, and "Unsure" to 2. The above transformation is applied to the training, development, and test datasets, ensuring consistency in label representation across all phases of model development.

3.3.2.2 Fine-tuning Encoder-based Transformer models. For this subtask, we explore various Encoder-based Transformer models to ensure precise and accurate predictions. Specifically, we fine-tuned BERT, BERT-Chinese, DistilBERT-Multilingual, DistilBERT-Uncased, and RoBERTa-Base using the official training set to adapt them to the characteristics of the dataset. The training effectiveness of each model was evaluated on the official development set, allowing for a comparative analysis of their performance in handling Chinese textual data and assessing claim validity duration.

All models were fine-tuned using their corresponding tokenizers for tokenization, ensuring proper text preprocessing and alignment with each model's architecture. We also experimented with different parameter values to analyze their impact on model performance, allowing us to identify the optimal model configuration.

For all models, the AdamW optimizer was utilized to enhance training stability and performance. The maximum sequence length was tested with values of 128 and 256 to assess the impact of input length on model effectiveness. The learning rate was explored with $2e-5$ and $1e-5$, while the batch size was varied across 32, 64, and 128 to evaluate its effect on training efficiency and convergence. Additionally, the number of epochs was tested with 3, 4, and 5 to

determine the optimal training duration for achieving the best model performance. An overview table summarizing the experimental settings and hyperparameter configurations is presented in Table 3.

Table 3: Fine-tuning Encoder-based Transformer Model Hyperparameter Range

| Hyperparameter | Value |
|----------------|-------------|
| Max Length | 128, 256 |
| Batch Size | 32, 64, 128 |
| Epochs | 3, 4, 5 |

3.3.2.3 Model Selection. After adjusting and testing various model parameters, we ultimately selected the two best-performing Encoder-based Transformer models for our final submission in this subtask: BERT-Chinese and DistilBERT-Multilingual. The detailed performance results are presented in section 4.

3.3.3 Fine-tuning GPT-4o-mini and Prompt-based Learning. In addition to evaluating BERT, an encoder-based Transformer model, we also examined GPT-4o-mini, a decoder-based model, to conduct a comparative analysis of their effectiveness in this task. This comparison provides insights into how different Transformer architectures influence model performance and task-specific adaptability.

3.3.3.1 Fine-tuning GPT 4o mini. When fine-tuning GPT-4o mini, we utilized the official training set and conducted the fine-tuning process through OpenAI's official platform. The development set was used to evaluate the model's performance after fine-tuning. The following presents the parameter configurations used in the fine-tuning process.

Table 4: Fine-tuning GPT 4o mini Hyperparameter Settings

| Hyperparameter | Value |
|----------------|-------|
| LR multiplier | 0.8 |
| Batch Size | 16 |
| Epochs | 3 |

3.3.3.2 Prompt-based Learning. In contrast to BERT, which directly processes raw text as input, GPT-4o Mini adopts a prompt-based approach, where the input is structured with an instructional prompt to guide the model toward the classification task. This method ensures that the model correctly interprets the context before generating a response.

The prompts used in GPT-4o Mini can be categorized into two types: System Prompt and User Prompt. The System Prompt defines the model's behavior and response style, providing overarching guidance to ensure consistent interpretation and processing of inputs. In contrast, the User Prompt consists of

specific queries or instructions given by the user, directly guiding the model to generate task-specific responses. The following are the System Prompt and User Prompt used in this subtask.

system_prompt =

*"You are a professional text classification assistant.
Please follow the rules below to classify the input text
and estimate the duration of its impact.
There are three classification labels for impact
duration:
1. Longer than 1 week
2. Within 1 week
3. Unsure
Please output only the final label without providing any
explanations.*

The User Prompt follows a few-shot learning approach, where examples from the dataset are provided to help the model learn and generate predictions based on the given patterns.

User Prompt =

*"The following are examples of texts along with their
corresponding label :*

example 1 :

*Text: 理論上今年可能還有開發金收購的動作,畢竟離
2020 初 100%持有已剩不到 2 年了. 近期觀察股價很硬,應該有人
持續在收籌碼等開發金動作...*

Label: Longer than 1 week

example 2 :

*Text: 其實 Band 1 也很多國家開放
嘿嘿嘿~~ 會不會在明年呢???*

Label: Longer than 1 week

example 3 :

*Text: 昨天真的是仙人指路的預告, 中毒繼續衝。
Label: Within 1 week*

*Please classify the following text based on the provided
classification rules:*

Text to classify:

[Text]

Respond only with one of the following labels:

- *Longer than 1 week*
- *Within 1 week*
- *Unsure*

The detailed GPT-4o Mini performance results are presented in Section 4.

3.4 Evaluation Metrics

For model evaluation, we use Micro-F1 and Macro-F1, which are the official ranking criteria of the competition. These metrics provide a balanced assessment of model performance by considering both overall accuracy and class-wise performance.

Micro-F1 calculates precision and recall across all instances before computing the F1-score. This approach gives equal weight to each instance, making it particularly effective in datasets with class imbalances, as it reflects the model's ability to correctly classify the majority of data points.

Macro-F1 computes the F1-score for each class separately and then averages the scores across all classes. Unlike Micro-F1, it treats all classes equally, regardless of their frequency in the dataset. This makes it useful for evaluating how well the model handles underrepresented categories, ensuring that performance is not dominated by the most frequent classes.

By considering both Micro-F1 and Macro-F1, we assess the model's overall effectiveness while ensuring fair evaluation across different class distributions.

4 EXPERIMENTS

4.1 Detection of Argument Temporal References (Earnings Conference Call)

This section presents the results of the argument temporal reference classification task. The models were evaluated based on fine-tuning and prompt-based learning strategies described in Chapter 3. Encoder-based transformer models, including RoBERTa-base and DistilBERT, were fine-tuned for classification, while GPT-4o mini was evaluated both before and after fine-tuning, using different prompting techniques.

4.1.1 Fine-tuning Encoder-based Transformer Models. To evaluate the effectiveness of encoder-based models, RoBERTa-base and DistilBERT-base were fine-tuned using a balanced training dataset. The hyperparameter configurations for each model are presented in Table 5.

Table 5: Hyperparameter Settings for Fine-Tuned Encoder-based Models

| Model | Learning rate | Max Length | Batch Size | Epochs |
|--------------------------------|---------------|------------|------------|--------|
| IMNTPU_ECC_1 (RoBERTa-base) | 5e-5 | 128 | 16 | 3 |
| IMNTPU_ECC_2 (DistilBERT-base) | 3e-5 | 128 | 16 | 5 |
| IMNTPU_ECC_3 (DistilBERT-base) | 1.5e-5 | 128 | 32 | 6 |

Each model was evaluated on both the validation dataset and the official test set. Table 6 presents the Micro-F1 and Macro-F1 scores for each model across both datasets.

Table 6: Performance of Fine-Tuned Encoder-based Models

| Model | Validation set | | Test set | |
|--------------------------------|----------------|----------|----------|----------|
| | Micro F1 | Macro F1 | Micro F1 | Macro F1 |
| IMNTPU_ECC_1 (RoBERTa-base) | 74.22% | 74.49% | 69.05% | 67.06% |
| IMNTPU_ECC_2 (DistilBERT-base) | 77.33% | 73.40% | 63.10% | 57.87% |
| IMNTPU_ECC_3 (DistilBERT-base) | 74.67% | 74.75% | 65.48% | 62.44% |

DistilBERT-based models achieved competitive performance, with IMNTPU_ECC_2 obtaining the highest Micro-F1 score of 77.33% on the validation dataset. RoBERTa-base also performed well, demonstrating that both architectures effectively capture temporal references in financial texts.

The performance of all models decreased on the official test set, with RoBERTa-base achieving the best result (Micro-F1: 69.05%). This suggests potential overfitting to the training dataset or differences in data distribution between the validation and official test sets.

4.1.2 Fine-tuning and Prompt-based Learning for GPT-4o Mini. GPT-4o Mini was evaluated using both fine-tuning and prompt-based learning. After fine-tuning, both pretrained and fine-tuned models were evaluated using zero-shot, one-shot, three-shot, and six-shot prompting strategies. The results on the validation dataset are shown in Table 7.

Table 7: Validation Set Results for GPT-4o Mini (Pretrained vs. Fine-tuned)

| Prompting Strategy | Pretrained GPT-4o Mini | | Fine-tuned GPT-4o Mini | |
|---------------------|------------------------|----------|------------------------|----------|
| | Micro F1 | Macro F1 | Micro F1 | Macro F1 |
| zero-shot learning | 65.87% | 60.83% | 65.87% | 62.94% |
| one-shot learning | 62.27% | 57.9% | 66.40% | 62.94% |
| three-shot learning | 65.35% | 60.60% | 67.73% | 64.32% |
| six-shot learning | 64.27% | 60.95% | 69.20% | 66.67% |

In the pretrained model, performance did not consistently improve as more examples were added in the prompt, suggesting that GPT-4o Mini already possessed a general understanding of the task but did not significantly benefit from additional context. However, after fine-tuning, the model demonstrated clear improvement, with Micro-F1 and Macro-F1 scores increasing progressively from zero-shot to six-shot learning, indicating that fine-tuning enhanced its ability to leverage in-context learning effectively.

4.1.3 Analysis of Results. Both encoder-based and decoder-based models demonstrated strong performance in temporal argument classification, but each has distinct strengths and limitations. RoBERTa-base and DistilBERT achieved high Micro-F1 scores on the validation dataset, confirming their ability to extract structured representations from financial text. However, their performance declined on the official test set, suggesting potential overfitting or domain differences between datasets.

GPT-4o Mini exhibited strong zero-shot capabilities, but its performance did not consistently improve with additional examples in the prompt, indicating that in-context learning alone was insufficient without fine-tuning. After fine-tuning, however, GPT-4o Mini showed significant gains, achieving its highest Micro-F1 score (69.20%) with six-shot prompting. This suggests that fine-tuning enhances the model’s ability to leverage in-context learning effectively.

A comparison between fine-tuned encoder-based models and GPT-4o Mini indicates that six-shot fine-tuned GPT-4o Mini performed comparably to the best encoder-based models on the validation dataset. These findings highlight the potential of combining encoder-based structured learning with decoder-based prompt-based learning for a more flexible and efficient approach to financial argument classification.

4.2 Assessment of the Claim's Validity Period (Social Media)

4.2.1 Result of Encoder-based Transformer Model. This section presents the final hyperparameter settings and results of the Encoder-based Transformer models. After testing and adjustments, the final hyperparameter settings used are summarized in Table 8.

Table 8: Hyperparameter setting

| Model | Hyperparameter | | |
|---|----------------|------------|--------|
| | Max Length | Batch Size | Epochs |
| IMNTPU_Social Media_2 (Bert Chinese) | 256 | 128 | 4 |
| IMNTPU_Social Media_3 (DistilBERT multilingual) | 256 | 128 | 4 |

We fine-tuned BERT, BERT-Chinese, DistilBERT-Multilingual, DistilBERT-Uncased, and RoBERTa-Base. Validation results on the development set indicate that these models achieve similar Micro F1 scores, ranging between 72% and 75%. However, in terms of Macro F1, there is a more noticeable performance gap. BERT-Chinese and DistilBERT-Multilingual achieved 57.3% and 56.93%, respectively, while the other models scored around 48%.

In the official Evaluation Results, the model performance was slightly lower compared to the validation results. The detailed results are presented in Table 9.

Table 9: performance of Encoder-based Transformer Model

| Model | Development Set Result | | Test Set Result | |
|---|------------------------|----------|-----------------|----------|
| | Micro F1 | Macro F1 | Micro F1 | Macro F1 |
| IMNTPU_Social Media_2 (Bert Chinese) | 75% | 57.3% | 72.83% | 53.40% |
| IMNTPU_Social Media_3 (DistilBERT multilingual) | 72.72% | 56.93% | 69.98% | 53.50% |

4.2.2 Result of Finetuned-GPT-4o mini

For the GPT model evaluation, we first tested the GPT-4o mini on the development set. Then, using the same prompt, we evaluated the fine-tuned GPT-4o mini. The results showed a significant performance gap between the two, with the fine-tuned model achieving performance comparable to the Encoder-based Transformer models.

Comparing the official results with the validation set results, the Micro F1 score in the official evaluation is higher than that of the validation set, while the Macro F1 is lower instead. The detailed results are presented in Table 10.

4.1.3 Analysis of Results. Overall, the experimental results demonstrate that both types of models performed well on this task. While GPT was not the best-performing model during the validation phase, it achieved the highest score in the final competition results. This suggests that large language models (LLMs) possess strong language comprehension capabilities, highlighting their potential for future advancements in Assessment of the Claim’s Validity Period.

Table 10: performance of GPT-4o mini

| Model | GPT-4o mini with Development set | | Finetuned GPT-4o mini with Development set | | Finetuned GPT-4o mini with Test set | |
|----------------------|----------------------------------|----------|--|----------|-------------------------------------|----------|
| | Micro F1 | Macro F1 | Micro F1 | Macro F1 | Micro F1 | Macro F1 |
| IMNTP Social Media_1 | 50.46 % | 47.54 % | 73.52 % | 57.45 % | 76.83 % | 54.68 % |

5 CONCLUSIONS

This paper outlines IMNTPU's participation in the NTCIR-18 FinArg-2 competition, focusing on two subtasks: Detection of Argument Temporal References and Assessment of the Claim's Validity Period. We employed data augmentation, fine-tuning, and prompt-based learning to evaluate the effectiveness of encoder-based and decoder-based Transformer models.

In Subtask ECCs, we addressed data imbalance using GPT-based data augmentation to create a more balanced dataset. Fine-tuning improved RoBERTa-base and DistilBERT's ability to detect temporal references, while GPT-4o Mini showed strong zero-shot capabilities but required fine-tuning for optimal prompt-based learning performance. Six-shot prompting on fine-tuned GPT-4o Mini achieved results comparable to the best encoder-based models.

In Subtask Social Media, encoder-based models performed well in validation, but fine-tuned GPT-4o Mini achieved the highest Micro-F1 score in the official evaluation, demonstrating the potential of large language models (LLMs) for financial text classification.

This study contributes to financial NLP research by demonstrating the effectiveness of integrating fine-tuning and prompt-based learning for financial argument classification. The comparative analysis of encoder-based and decoder-based models provides insights into their respective strengths, with encoder-based models excelling in structured classification and decoder-based models benefiting from prompt engineering. The results underscore the potential of combining both approaches to enhance financial text analysis.

From a managerial perspective, the findings highlight the practical benefits of leveraging LLMs for financial text classification. Organizations and financial analysts can adopt fine-tuned generative models to improve text-based decision-making while reducing dependence on extensive labeled datasets. Additionally, the study suggests that integrating prompt engineering strategies can enhance the adaptability of financial AI applications to dynamic market conditions.

Future work will explore more effective collaboration between different models to enhance financial text classification. This includes integrating multiple generative models, where each specializes in distinct aspects of financial reasoning, such as numerical analysis, sentiment interpretation, or contextual inference. Further research may also focus on adaptive learning mechanisms, allowing models to refine their understanding of evolving financial trends without frequent retraining.

ACKNOWLEDGMENT

This work was supported by National Science and Technology Council, Taiwan, under grants NSTC 113-2425-H-305-003-, 114-2425-H-305-003-, National Taipei University (NTPU), Taiwan and ATEC Group under grants NTPU-112A413E01, and National Taipei University (NTPU), Taiwan under grants 114-NTPU_ORDA-F-004.

REFERENCES

- [1] A. Alhamzeh, "Financial argument quality assessment in earnings conference calls," in *International Conference on Database and Expert Systems Applications*, 2023: Springer, pp. 65-81.
- [2] C. Chiu Jr, C.-C. Chen, H.-H. Huang, and H.-H. Chen, "Pre-Finetuning with Impact Duration Awareness for Stock Movement Prediction," *arXiv preprint arXiv:2409.17419*, 2024.
- [3] J. Lee and K. Toutanova, "Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, vol. 3, no. 8, 2018.
- [4] M. V. Koroteev, "BERT: a review of applications in natural language processing and understanding," *arXiv preprint arXiv:2103.11943*, 2021.
- [5] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: a hybrid model for sentiment analysis with transformer and recurrent neural network," *IEEE Access*, vol. 10, pp. 21517-21525, 2022.
- [6] J. Achiam *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
- [7] Y. Peskine, D. Korenčić, I. Grubisic, P. Papotti, R. Troncy, and P. Rosso, "Definitions Matter: Guiding GPT for Multi-label Classification," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 4054-4063.
- [8] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacl-HLT*, 2019, vol. 1: Minneapolis, Minnesota, p. 2.
- [9] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification?," in *Chinese computational linguistics: 18th China national conference, CCL 2019, Kunming, China, October 18–20, 2019, proceedings 18*, 2019: Springer, pp. 194-206.
- [10] C. W. Mayer, S. Ludwig, and S. Brandt, "Prompt text classifications with transformer models! An exemplary introduction to prompt-based learning with large language models," *Journal of Research on Technology in Education*, vol. 55, no. 1, pp. 125-141, 2023.
- [11] L. Loukas, I. Stogiannidis, P. Malakasiotis, and S. Vassos, "Breaking the bank with ChatGPT: few-shot text classification for finance," *arXiv preprint arXiv:2308.14634*, 2023.
- [12] J. White *et al.*, "A prompt pattern catalog to enhance prompt engineering with chatgpt," *arXiv preprint arXiv:2302.11382*, 2023.
- [13] J. Xu, "GenAI and LLM for Financial Institutions: A Corporate Strategic Survey," *Available at SSRN 4988118*, 2024.
- [14] J. Wei and K. Zou, "Eda: Easy data augmentation techniques for boosting performance on text classification tasks," *arXiv preprint arXiv:1901.11196*, 2019.
- [15] J. M. Tapia-Télez and H. J. Escalante, "Data augmentation with transformers for text classification," in *Mexican International Conference on Artificial Intelligence*, 2020: Springer, pp. 247-259.
- [16] L. Vogel and L. Flek, "Investigating Paraphrasing-Based Data Augmentation for Task-Oriented Dialogue Systems," in *International Conference on Text, Speech, and Dialogue*, 2022: Springer, pp. 476-488.
- [17] C. Suhaeni and H.-S. Yong, "Mitigating Class Imbalance in Sentiment Analysis through GPT-3-Generated Synthetic Sentences," *Applied Sciences*, vol. 13, no. 17, p. 9766, 2023.
- [18] X.-S. Hong, S. Wu, M. Tian, and J. Jiang, "CYUT at the NTCIR-16 FinNum-3 Task: Data Resampling and Data Augmentation by Generation," in *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo Japan, 2022*, pp. 14-17.
- [19] Y. Rebboud, P. Lisena, and R. Troncy, "Prompt-based Data Augmentation for Semantically-Precise Event Relation Classification," in *SEMMES 2023*, 2023.
- [20] F. Sufi, "Generative pre-trained transformer (GPT) in research: A systematic review on data augmentation," *Information*, vol. 15, no. 2, p. 99, 2024.
- [21] C.-C. Chen *et al.*, "Overview of the NTCIR-18 FinArg-2 Task: Temporal Inference of Financial Arguments," presented at the Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo, Japan, 2025, 2025.