

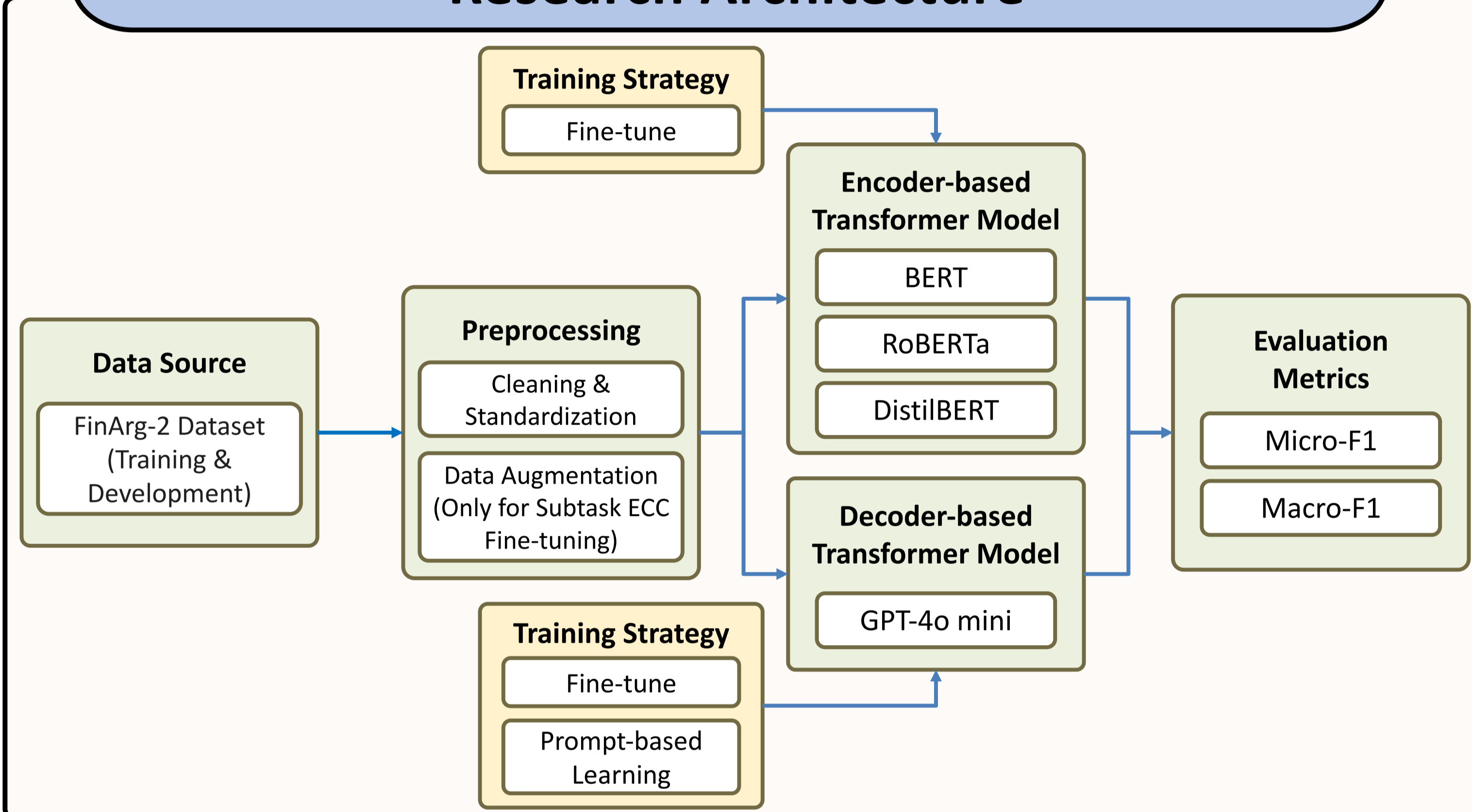
Bor-Jen Chen¹Wen-Hsin Hsiao¹Jun-Yu Wu²Cheng-Yun Wu¹Min-Yuh Day^{1, *}

*myday@gm.ntpu.edu.tw

¹Information Management, National Taipei University, New Taipei City, Taiwan²Leisure and Sport Management, National Taipei University, New Taipei City, Taiwan

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.

Research Architecture



Performance

Performance of Fine-Tuned Encoder-based Models on ECCs

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%

Validation Set Results for GPT-4o mini on ECCs

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%

Performance of GPT-4o mini on Social Media

Model	GPT-4o mini (Development)		Finetuned GPT-4o mini (Development)		Finetuned GPT-4o mini (Test)	
	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1
IMNTPU Social Media 1 (Finetuned GPT-4o mini)	50.46%	47.54%	73.52%	57.45%	76.83%	54.68%

Performance of Fine-Tuned Encoder-based Models on Social Media

Model	Development Set		Test Set	
	Micro F1	Macro F1	Micro F1	Macro F1
IMNTPU Social Media 2 (Bert Chinese)	75%	57.3%	72.83%	53.40%
IMNTPU Social Media3 (DistilBERT multilingual)	72.72%	56.93%	69.98%	53.50%

Conclusions

- Fine-tuned RoBERTa (Micro-F1: 69.05%) and DistilBERT (Micro-F1: 65.48%) performed well on ECCs
- 6-shot GPT-4o mini (Micro-F1: 69.20%) matched top encoder models on ECCs
- Fine-tuned GPT-4o mini achieved the highest Micro-F1 in the official test on Social Media
- Encoder models were stable; decoder models improved significantly with prompt engineering on Social Media
- Fine-tuning + prompt engineering boosts performance and improves adaptability in financial

Acknowledgement

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.

ECCs Experiments

Fine-tuned Encoder-based Model Hyperparameter setting:	Fine-tuned GPT 4o-mini Hyperparameter setting:
<ul style="list-style-type: none"> Max Length : 128 Batch Size : 16 / 16 / 32 Epochs : 3 / 5 / 6 Learning rate: 5e-5 / 3e-5 / 1.5e-5 	<ul style="list-style-type: none"> Learning rate: 0.1 Batch Size : 4 Epochs : 3

Prompt Design

Zero-shot 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: [Claim]

Premises: [Premises]

Label: [Label]

In one-shot, three-shot, and six-shot learning, the prompt was extended with labeled examples before presenting the test instance.

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

""The number of examples varies depending on the number of shots.""

Social Media Experiments

Fine-tuned Encoder-based Model Hyperparameter setting:	Fine-tuned GPT 4o-mini Hyperparameter setting:
<ul style="list-style-type: none"> Max Length : 256 Batch Size : 128 Epochs : 4 	<ul style="list-style-type: none"> Learning rate: 0.8 Batch Size : 16 Epochs : 3

Prompt Design

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:

- Longer than 1 week
- Within 1 week
- Unsure

Please output only the final label without providing any explanations.

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,應該有人持續收籌碼等開發金動作...

(In theory, there might still be acquisition moves by KGI this year — after all, it's less than two years away from full ownership since early 2020. Recently, the stock price has been holding strong, suggesting that someone may be accumulating shares in anticipation of KGI's next move.)

Label: Longer than 1 week

Example 2 :

Text: 其實Band 1也很多國家開放 嘿嘿~ 會不會在明年呢??? (Actually, many countries have already opened up Band 1 — hehe~ could it be our turn next year???)

Label: Longer than 1 week

Example 3 :

Text: 昨天真的是仙人指路的預告, 中壽繼續衝 (Yesterday was truly a prophetic sign — Chung Shou keeps on soaring!)

Label: Within 1 week

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

Text to classify: [Text]