Taiwan Scam Detect on Threads

Modeling & Evaluating

Natural Language Processing Group 8

Yan-Ru Liu and Che-Kuan Shen

Key words: AI-generated content (AIGC), large language model (LLM), Agentic AI, meta Llama, TAIDE

1 Introduction

The rise of online scams has become a pressing issue in Taiwan, reflecting global trends in digital fraud. With the increasing use of social media platforms, particularly text-based ones like Threads, scammers have discovered new avenues to exploit unsuspecting users. This study focuses on the detection and evaluation of scams on Threads, a burgeoning platform for community engagement and dialogue.

Our research aims to explore advanced modeling techniques to identify patterns and anomalies indicative of fraudulent activities. By leveraging data-driven approaches, such as machine learning and natural language processing (NLP), this study evaluates the effectiveness of various detection models in combating scams on Threads. The goal is to provide insights into the evolving landscape of scams and propose robust strategies for early detection, thereby enhancing platform security and user trust.

Through the integration of real-world data and evaluation metrics, this research contributes to a deeper understanding of scam dynamics in Taiwan and informs broader strategies for digital fraud prevention. It addresses a critical need to adapt to new challenges posed by the digital economy while safeguarding the integrity of online communities.

2 Methodology

To investigate effective strategies for scam detection on Threads, we designed and implemented three experimental methodologies leveraging state-of-the-art AI models and techniques. Each approach was tailored to perform the task of text classification, focusing on identifying fraudulent content. The models and tools used in the experiments include the taide/Llama3-TAIDE-LX-8B-Chat-Alpha1, MediaTek-Research/Breeze-7B-Instruct-v1_0 and Qwen2.5-1.5B-Instruct, with optimization parameters carefully selected for optimal performance.

2.1 Few-Shot Learning with Generative AI

In this approach, we utilized a generative AI model with few-shot learning capabilities. Carefully crafted prompts were used to instruct the model on identifying potential scam patterns based on limited labeled examples. This method evaluated the adaptability of pretrained language models to text classification tasks in low-resource settings.

2.2 Fine-Tuning Large Language Models

The second methodology involved fine-tuning the Taide/Llama3-TAIDE-LX-8B-Chat-Alpha1, MediaTek-Research/Breeze-7B-Instruct -v1_0 and Qwen/Qwen2.5-1.5B-Instruct models using a curated dataset of 13,510 training examples for scam detection. The fine-tuning process utilized the **adamw_torch** optimizer with a learning rate of **1e-4**. This setup aimed to enhance the models' contextual understanding and classification accuracy by adapting them to domain-specific data.

2.3 Agentic AI Implementation with LangGraph

In our third experimental approach, we implemented an agentic AI framework for scam detection on Threads, drawing inspiration from recent advancements in retrieval-augmented generation (RAG) models. This method integrates retrieval mechanisms with large language models (LLMs) to enhance the accuracy and robustness of text classification tasks. We utilized the LangGraph framework to develop an autonomous AI agent capable of dynamically retrieving relevant information and generating responses based on the retrieved data. This design allows the agent to adapt its retrieval and generation strategies according to the complexity of the input queries, ensuring efficient handling of both straightforward and complex cases.

Our approach, as depicted in Figure 1, is informed by methodologies such as Adaptive-RAG, which selects appropriate strategies based on query complexity, and Self-RAG, which enhances generation quality through self-reflection and critique. By incorporating these principles, our agentic AI system aims to improve the detection of scam-related content on Threads, offering a more responsive and context-aware solution compared to traditional methods.

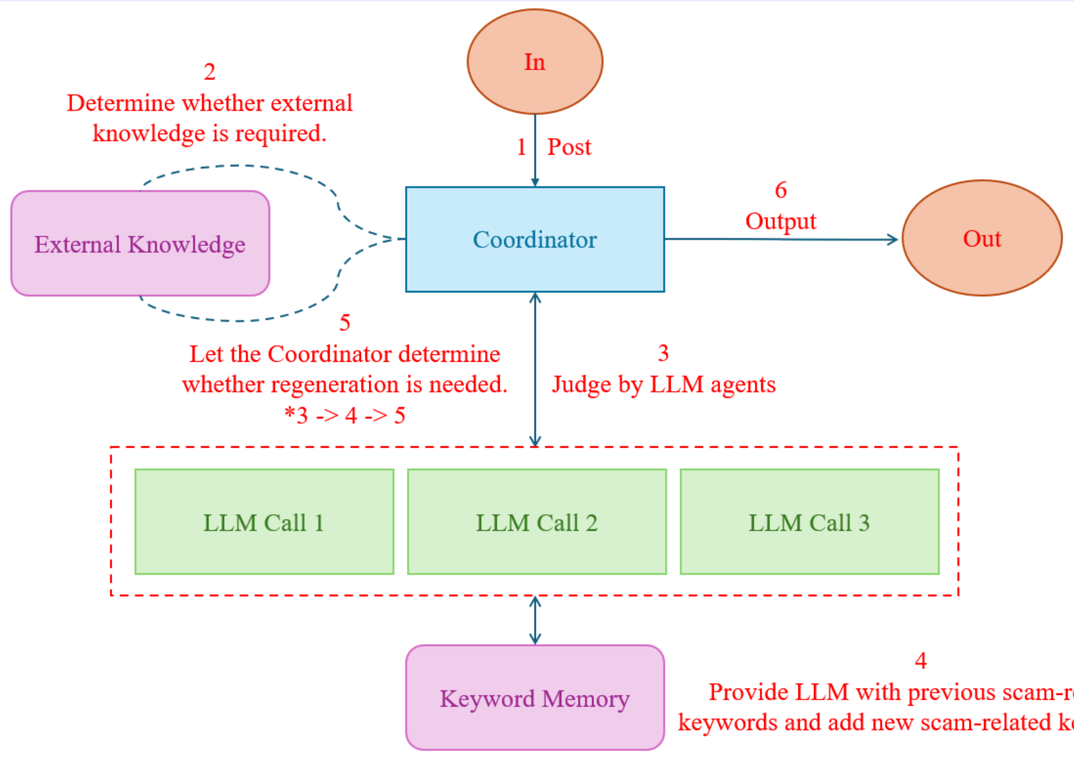

Figure 1 : **Workflow of the Agentic AI**. The diagram illustrates the operation of Agentic AI designed for detecting and addressing scam-related post on threads.

3 Experiment results

We use F-1 score as our evaluation metric. Figure 2 presented that we compare methods across work, gambling, dating, and investment. Fine-tuned Llama3-TAIDE-LX-8B-Chat-Alpha1 outperforms other models in most categories, particularly in work and dating, suggesting that its higher parameter count and targeted training yield stronger domain comprehension. Yet, few-shot Llama3-TAIDE-LX-8B-Chat-Alpha1 excels in gambling, indicating that this approach might better capture certain patterns without extensive retraining. For investment, all models perform poorly, presumably due to the domain's complexity and limited relevant data; Breeze-7B-Instruct-v1_0 with few-shot learning shows a slight edge, though improvements are needed across the board. These trends underscore how both model capacity and training strategies must be aligned with category-specific demands.

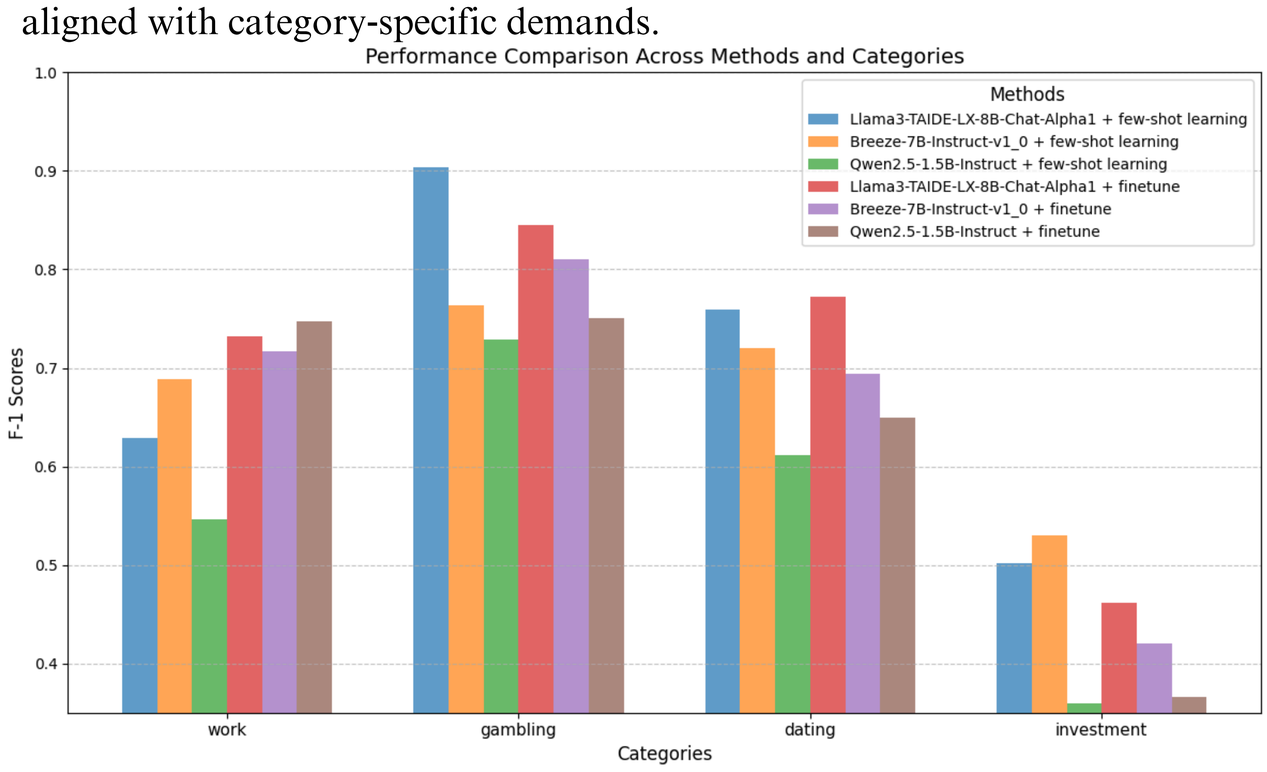

Figure 2 : **Performance Comparison Across Methods and Categories.** The chart illustrates the F1-scores of six different methods across four categories, with the following data distribution: work (750 entries), gambling (452 entries), dating (777 entries), and investment (963 entries).

4 Conclusion and Future Work

This study demonstrates the potential of AI-driven methods, including few-shot learning, fine-tuning, and agentic AI, in detecting scams on platforms like Threads. Fine-tuning showed better performance in most scenarios, while few-shot learning offered flexibility in specific cases. **Unfortunately, the development of agentic AI was incomplete due to time constraints but will be further refined in future work.**

Future work will focus on optimizing training methods, integrating dynamic real-world data, and enhancing domain-specific features. Expanding to other platforms and refining agentic AI approaches will further improve scam detection and user trust.