Taiwan Scam Detect on Threads
Data collection & Preprocessing

Natural Language Processing　Group 9
Hsin Ti Kuo，Zong-Yu Wu and Guang-Shuo Hsu

I Introduction

With the proliferation of social media platforms, fraudulent activities have become increasingly rampant worldwide, and Taiwan is no exception. Scammers exploit these platforms to disseminate false information, luring users into illegal transactions or divulging personal information. To effectively detect and prevent such fraudulent activities, our project focuses on the application of Natural Language Processing (NLP) techniques in detecting scam posts on social media.

This project is a collaborative effort between two groups. Our group is primarily responsible for detecting scam clues in Taiwan, including data collection and preprocessing. We have gathered a substantial amount of data across categories such as employment, gambling, relationships, and investments. Utilizing APIs, we have extracted relevant information, including post content, author IDs, number of responses, number of likes, and post IDs. Additionally, we have downloaded images associated with the posts for use by the model development team. Each post has been annotated to indicate whether it is fraudulent, thereby creating a comprehensive dataset.

To evaluate the model's performance, we have also collected tens of thousands of unannotated and uncleaned data points. These data will be used for model training and validation to enhance its accuracy and reliability in real-world applications.

ime

II Data Collection

In this study, we conducted two main phases of data collection:

1. Initial Search Using Scrapfly API

   We utilized the Scrapfly API to search for fraudulent posts on Threads. Before performing web crawling, we conducted preliminary filtering using keywords. Targeted webpages were then crawled to collect textual and related information. In this phase of data collection, the following items were obtained :

   Text, Text ID, User profile pictures, Number of replies, Number of likes, Image URLs

2. Large-Scale Data Collection

   To meet the requirements for model training and testing, we conducted additional large-scale data collection. A total of 20,784 unlabeled posts were collected, not using the Scrapfly API but through custom web scraping scripts.

   The categories and their subtypes are as follows:

a. Job Scams (5,773 posts):

◦ make money easily	| 輕鬆賺錢
◦ Chatting for Money	| 聊天賺錢
◦ Make money with me	| 賺錢找我
◦ Make money by commenting	| 聊天賺錢
◦ Spend 5-10 minutes a day to earn 2000+	| 每天花5-10分鐘就能賺2000+

b. Gambling Scams (4,926 posts):

◦ Online Casinos/Baccarat	| 娛樂城/百家樂
◦ Sports Betting Analysis	| 運彩分析
◦ Free Credits/Top-ups	| 贈送儲值金
◦ High Return Rates	| 高倍返利
◦ Free communication group	| 免費交流群

c. Investment Scams (5,104 posts):

◦ Make money investing in cryptocurrencies	| 投資加密貨幣賺錢
◦ Zero-Risk Investment, Guaranteed Profit	| 零風險投資、穩賺不賠
◦ Limited time investment course discounts	| 投資課程限時優惠
◦ Investment advice for newbies	| 新手投資推薦
◦ passive income	| 被動收入

d. Pornographic Scams (4,981 posts):

◦ Seeking Boyfriend/Girlfriend	| 徵男友、徵女友
◦ Hookup Tools, Contact Me for Intimacy	| 約炮神器、做愛找我
◦ Paid Companionship Services	| 援交服務
◦ Adult Videos	| 成人影片
◦ Hostess Agent	| 酒店經濟

During the collection process, we categorized data into different fraud types using specific keywords to ensure high relevance.

III Data Preprocessing

During the data preprocessing stage, we handled both text and images with the following steps:

1. Emoji Conversion

   Converted emojis in the text into their corresponding textual descriptions to allow the model to better interpret and process the data.

2. Special Character Removal

   Removed all punctuation marks, underscores, ellipses (both full-width and half-width), special characters (such as hearts), and newline symbols. Only Chinese, English, and numerical characters were retained to ensure the text is clean and structured.

3. Image Storage Processing

   Images in the crawled data were initially stored as temporary links. To enable permanent usage by the model team, we downloaded the images and saved them locally. Simultaneously, we updated the image URLs in the JSON file to point to the local storage paths.

IV Logic of manual annotation

- Work Scams: Posts mentioning easy money-making methods, specific earnings, or cash are classified as scams. Posts with keywords but unrelated content are marked as non-scams.
- Gambling Scams: Posts offering free credits, doubled returns, or using AI predictions are scams. Pure recommendations or analyses without inducement are non-scams.
- Investment Scams: Detailed analyses with group invitations or links are labeled as scams. Posts sharing experiences with "for reference only" disclaimers are non-scams.
- Pornographic Scams: Posts with dating conditions, invitations, or explicit videos are scams. Posts using trending keywords without relevant content are marked as non-scams.

V Experiment results

These steps ensure the stability and consistency of the dataset, laying a solid foundation for model training. After large-scale data comparison and removal of duplicate texts, 13,510 unique entries remain. Table 1 illustrates the results of our data annotation process.

	Total	Scam	Non-Scam
Manual Annotation	3126	1102	2024
Language Model Annotation	13510	6414	7096

Table 1 : Data Quantity

V Conclusion

I This project focuses on collecting and analyzing various types of fraud-related data in Taiwan to establish a dataset specifically targeting fraudulent posts on Threads. We hope that by integrating this data with natural language processing techniques, it can provide effective technical support for detecting and preventing social media fraud. Additionally, we aim to enhance public awareness and recognition of fraudulent activities, ultimately fostering a safer and more harmonious online environment.

Tools : Scrapfly Web Scraping API https://scrapfly.i