LLM Agent for Disinformation Detection Based on DISARM Framework

Kevin Tseng, Man-Kwan Shan
National Chengchi University
111971016@g.nccu.edu.tw, mkshan@nccu.edu.tw

Abstract

The rise of AI-generated disinformation driven by large language models (LLMs) exacerbates threats to information trust and democracy. Motivated by the need for efficient detection, this paper integrates the DISARM framework with LLMs to enhance capabilities. The proposed OSINT LLM Agent leverages DISARM's structure and LLMs' reasoning for an effective, adaptable solution against disinformation campaigns. Experiments using real-world cases demonstrate feasibility and effectiveness, highlighting DISARM-LLM integration's potential for advanced detection.

Introduction

The rapid spread of disinformation campaigns in the digital age poses significant threats to trust in information sources and democratic institutions worldwide. The investigation of disinformation over social media often relies on open-source intelligence (**OSINT**) techniques. OSINT involves collecting and analyzing publicly available information from various sources, such as social media platforms, news websites, and online databases for intelligence purposes to identify and counter disinformation campaigns.

However, traditional manual investigations struggle to keep pace with the scale and complexity of disinformation spreads over social media. Recent advancements in Large Language Models (LLMs) offer promising opportunities to enhance disinformation detection capabilities and streamline OSINT processes. In particular, LLMs have shown remarkable potential in task automation and natural language processing (Shafee, Bessani, and Ferreira 2024), and techniques like Chain of Thought (**CoT**) reasoning (Wei et al., 2022) and Reason+Act (**ReAct**, Yao et al., 2022) have further enhanced their reasoning capabilities and interaction with external resources. While CoT guides LLMs to deal with hard problems by breaking into step-by-step reasoning process, ReAct combines reasoning and acting in an interleaved manner to allow LLMs to interact with external resources.

This paper proposed the OSINT LLM agent based on CoT and ReAct for the investigation and detection of disinformation spreads over social media. To accommodate the concept of CoT and ReAct, we proposed to develop the OSINT LLM agent by integrating with the DISARM framework.

The **DISARM** (Disinformation Analysis and Response Measures) framework, proposed by the non-partisan, non-profit DISARM Foundation, aims to provide a systematic methodology for describing and understanding disinformation incidents (Terp and Breuer 2022). **DISARM** [1]organizes disinformation campaigns into four phases: Plan, Prepare, Execute, and Assess. Each phase encompasses various tactics and techniques employed by adversaries. For instance, the "Prepare" phase includes tactics like "Develop Content," which involves creating or acquiring text, images, and other content.

DISARM also proposes detection methods to identify and analyze disinformation activities. For example, the detection method (F00015) "Detect Anomalies in Membership Growth Patterns," by monitoring unusual growth patterns in social media group memberships or online funding campaigns, can help to identify techniques such as "Create Inauthentic Social Media Pages and Groups" (T0007), "Create Fake Experts" (T0009), and "Organize Events" (T0057).

This paper presents a novel approach integrating the DISARM framework with LLMs to improve disinformation detection. By leveraging LLMs' capabilities within DISARM's systematic structure, we aim to develop a more effective solution for detecting and mitigating disinformation campaigns, enhancing the efficiency of OSINT investigations.

Related Work

Disinformation detection primarily relies on manual investigations with limited AI assistance. LLMs have shown potential in automating tasks and handling NLP tasks (Shafee, Bessani, and Ferreira 2024). Recent advancements, such as CoT reasoning (Wei et al., 2022) and ReAct (Yao et al., 2022), enhance LLMs' reasoning and interaction capabilities,

[1] https://disarmframework.herokuapp.com/

showing promise in improving the effectiveness of LLMs in detecting and analyzing disinformation.

Several frameworks have been proposed to analyze disinformation. Ben Nimmo's 4D model (Nimmo 2015) describes disinformation tactics as Dismiss, Distort, Distract, and Dismay. Another approach is to adapt models from cybersecurity, such as the Cyber Kill Chain (Hutchins, Cloppert, and Amin, 2011), which outlines the stages of a cyberattack. Disinformation can be seen as a form of cyberattack, making the Cyber Kill Chain a relevant model for understanding disinformation campaigns. The DISARM framework (Terp and Breuer, 2022) builds upon these concepts, providing a structured approach to analyzing disinformation. This paper introduces an OSINT LLM Agent that integrates the DISARM framework with LLMs for better disinformation detection.

Method

The proposed OSINT LLM Agent integrates the DISARM framework with LLMs through five interconnected modules. First, the Data Collection Tool Selection module identifies suitable tools for gathering relevant data. Next, the Data Processing module transforms raw data into a structured format for LLM analysis. Moving forward, the Analysis Strategy Proposal module devises targeted approaches based on DISARM's detection methods. Subsequently, the Autonomous Analysis Execution module leverages LLMs to perform analysis strategies, generating insights. Finally, the DISARM TTPs Identification module pinpoints the specific disinformation campaign, updating the knowledge base. The modular architecture ensures flexibility and adaptability across a wide range of disinformation detection scenarios. Figure 1 shows the architecture of proposed LLM agent.

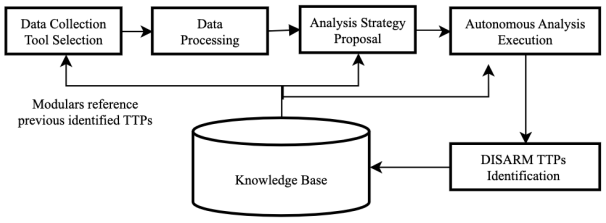

Figure 1: OSINT LLM Agent Workflow.

While these five modules do not strictly adhere to the definition of CoT reasoning, they share a similar principle of breaking down complex reasoning processes into smaller, more manageable sub-tasks, which is crucial for effectively addressing the challenges associated with disinformation detection.

The specific LLMs, techniques, and skills used in our method are designed to be replaceable and adaptable based on the task requirements and computation resources available. For instance, while we employ ReAct in our experiments, it can be seamlessly replaced by alternative techniques that enable the LLM to reason, act, and interact with the environment. This replaceability allows our method to be compatible with a wide range of LLMs, techniques, and skills, making it a versatile solution for tackling evolving disinformation challenges.

It is worth noting that OSINT encompasses a wide range of scenarios, such as keyword searching, actor analysis, and social network analysis. The process typically involves narrowing down from a broad scope to specific actors, iteratively confirming anomalous groups, and investigating their tactics and intentions. Fully automating these complex tasks remains challenging at the current stage and requires further experimentation and exploration. The experiment presented in this paper simulates a relatively simple scenario of keyword searching on YouTube to discover patterns. However, the results demonstrate the promising future potential of the OSINT LLM Agent in more complex and diverse OSINT scenarios.

By decomposing the task of disinformation detection into five distinct stages, our method enables the LLM agent to analyze data, generate insights, and identify **TTPs** (Tactics, Techniques, and Procedures), which refer to the patterns of activities and methods employed by threat actors to achieve their goals. In this context, TTPs are based on the DISARM framework's taxonomy, which categorizes the various tactics and techniques used in disinformation campaigns. Our method employs a structured and systematic approach, utilizing various techniques such as CoT-inspired reasoning, to identify and classify these TTPs.

(1) Data Collection Tool Selection The agent selects the most appropriate data collection tool based on the previously identified DISARM TTPs stored in the knowledge base. The choice of tool depends on the investigation's requirements and the data sources.

Data can be collected from various sources, including search engines (Google, Bing), social media platforms (TikTok, YouTube), and other relevant websites or databases. In cases where the data sources are targeted at specific platforms or websites, the agent may prioritize the use of tools designed for those platforms. For example, if the investigation primarily involves data from social media platforms like Twitter or Facebook, the agent may opt for tools specifically developed for scraping or analyzing data from these platforms. The agent can choose from a variety of data collection methods, such as utilizing available APIs, employing custom data collection programs, or leveraging existing tools designed for specific platforms. The selection of appropriate data collection approach is crucial, as it directly

impacts the quality and relevance of the data gathered for the investigation.

The data collection process plays a pivotal role in the OSINT LLM Agent's effectiveness, as the accuracy and reliability of the subsequent analysis heavily rely on the quality and relevance of the collected data. By carefully selecting the appropriate data collection tools and sources and preprocessing the collected data into a structured format, the agent can optimize the disinformation detection process, leading to more accurate insights and enhanced overall performance of the method.

(2) Data Processing The agent processes and structures the collected data into tabular format by sampling due to LLM token limits. This organizes the data for the LLM to analyze and generate insights. Structured tabular data enables efficient data manipulation, filtering, and aggregation to detect disinformation patterns.

(3) Analysis Strategy Proposal The agent considers potential analysis approaches and outlines detailed execution strategies based on sample data and DISARM framework's detection methods. To ensure targeted analysis, the strategy proposal is divided into two stages:

(3.1) Column Selection This step focuses the information by concentrating on key columns, reducing the scope of consideration for the agent. It helps prioritize the most relevant information, especially when dealing with large datasets.

(3.2) Propose Potential Analysis Strategy The agent proposes feasible analysis strategies based on the selected columns and detection methods, guided by the TTPs identified in previous iterations of the investigation process. This two-stage approach allows the agent to systematically focus the investigation, optimize the analysis, and improve the overall performance in detecting disinformation campaigns.

(4) Autonomous Analysis Execution We explored two approaches for the autonomous execution of the proposed analysis strategies, utilizing different language models within a REPL (Read-Eval-Print Loop) environment:

(4.1) OpenAI's GPT-3.5 model with functions-agent This approach uses the GPT-3.5-turbo-0125 model to generate and execute Python code. The model is provided with the necessary instructions and templates to perform the analysis.

(4.2) Anthropic's Claude-3 model in ReAct mode This approach employs the Claude-3-haiku-20240307 model, which continuously performs actions, observes outputs, and determines the next steps. The model is given instructions using the ReAct prompt template, enabling it to generate and execute code tailored to the analysis requirements.

The selection of model and approach is based on the investigation's requirements, analysis complexity, and LLM capabilities. The agent provides insights, identifying patterns, or indicators of disinformation, which form the basis for the subsequent steps in the OSINT LLM Agent's process.

(5) DISARM TTPs Identification and Knowledge Base Update The agent identifies and stores the TTPs employed in the disinformation campaign based on the DISARM framework. The modular OSINT LLM Agent combines DISARM, LLMs, and a knowledge base to streamline detection and refine capabilities, addressing various disinformation detection scenarios.

Experiments

To demonstrate the feasibility and effectiveness of our proposed OSINT LLM Agent, we conducted experiments using a real-world disinformation campaign. The campaign, dubbed "Secret History of Tsai Ing-wen(蔡英文秘史)," targeted the January 2024 Taiwan presidential election. The disinformation spread rapidly on YouTube, with AI-generated videos (Microsoft Threat Intelligence 2024).

In our experiments, we prioritized feasibility and cost-effectiveness by using readily available APIs, such as GPT-3.5 and Claude-3 Haiku. Users can adjust their choice of models based on specific needs, opting for more powerful models when complex reasoning is required.

(1) Data Collection Tool Selection In this experiment, we manually specified the YouTube tool for data collection and collected video metadata using keyword searches. The agent did not reference any previously identified DISARM TTPs, as it was the first round of searching.

(2) Data Processing The collected data was processed to create a sample dataset. This step, which does not involve LLM usage, is necessary to the token limits of the LLMs.

(3) Analysis Strategy Proposal The LLM agent considered potential analysis approaches and outlined execution strategies based on the sample data and DISARM framework's detection methods. To avoid hallucinations due to the large number of detection methods, we selected "Detect abnormal amplification" and "Detect anomalous activity" as substitutes.

(3.1) Column Selection This step focuses on key columns, reducing the scope for the agent. We use a prompt with two variables: **{detect_methods}** for selected detection methods and **{sample_rec}** for sample data, both dynamically inserted. Figure 2 gives the prompt of column selection.

Prompt

You are an OSINT Expert tasked with identifying critical data columns for the detection methods **{detect_methods}** within a dataset. Utilizing your expertise in analyzing publicly available data, review the sample data provided: **{sample_rec}**. For each important column, explain its significance in relation to the detection methods in a JSON format, where the key is the column name and the value is your expert rationale.

Figure 2. Prompt of Column Selection.

An example of the output result for one of **{detection_cols}** is as follows: {"view_count": "The view count column can help detect abnormal amplification by identifying videos that have an unusually high number of views compared to similar videos in the dataset. A sudden spike in view count may indicate artificial inflation of views. "}

(3.2) Propose Potential Analysis Strategy In the second step, based on the columns obtained in the previous step and the detection methods, please propose viable analysis strategies. Figure 3 gives the prompt proposing analysis strategy.

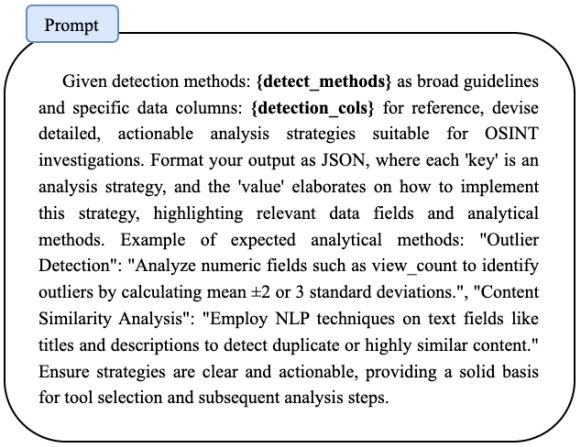

Figure 3. Prompt of Propose Potential Analysis Strategy.

One of the resulting strategies is: "Temporal Analysis": "Analyze the upload_date and view_count trends over time to identify sudden changes or irregular patterns. Look for videos with abnormal view_count growth rates or inconsistent upload_date patterns as potential indicators of abnormal amplification or anomalous activity."

(4) Autonomous Analysis Execution We explored two approaches for the autonomous execution of the proposed analysis strategies. Ideally, the LLM could be employed to choose the analysis method based on custom criteria or different execution strategies. Both models use the same **instruction**, selected from the analysis strategies. Figure 4 gives the instruction of autonomous analysis execution.

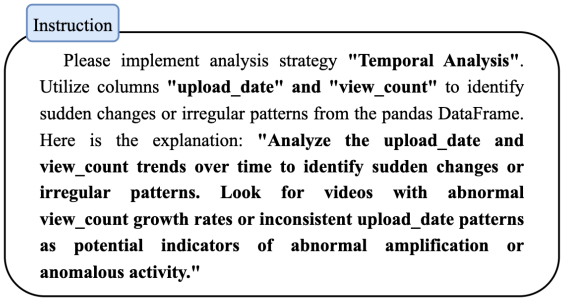

Figure 4. Instruction of Autonomous Analysis Execution.

(4.1) OpenAI's functions-agent by GPT-3.5 The GPT-3.5-turbo-0125 model generated Python code and ran it in a REPL (Read-Eval-Print Loop) environment. The prompt template used is imported from LangChain Hub using "langchain-ai/openai-functions-template" and combined with the instruction. Although the model claimed to have created visualizations using matplotlib and prompted the user to observe the graphs for insights, the actual output did not display any usable results. This suggests that the model was somewhat lacking in providing concrete, accessible findings.

The final answer provided by the model: I have plotted the "View Count Trends Over Time" based on the "upload_date" and "view_count" columns. This visualization will help us identify any sudden changes or irregular patterns in the view count over time.

(4.2) Anthropic's Claude3 model in ReAct mode The Claude-3-haiku-20240307 model continuously performed actions, observed outputs, and determined the next steps. The ReAct prompt template is imported from LangChain Hub using "langchain-ai/react-agent-template" and combined with the instruction. This approach yielded clear trend analysis results, with the agent generating and executing Python code tailored to the analysis requirements, such as conducting time-series analysis on the upload_date and view_count fields to identify anomalous patterns.

The brief final answer provided by the model: The temporal analysis of the "upload_date" and "view_count" columns identified videos with sudden changes and irregular patterns in view count growth rates. These videos, such as Abcd1234 and Abcd1235 (hashed output id), may indicate abnormal amplification, anomalous activity, or potential manipulation, warranting further investigation.

Conclusions

This paper proposed the LLM agent for disinformation detection over social media based on the DISARM framework. The OSINT LLM Agent leverages both frameworks, demonstrating adaptable solution potential. Pilot experiment was performed on the disinformation detection of Taiwan's President Tsai Ing-wen over YouTube during the 2024 presidential election. Future work includes of enabling multi-round investigations, testing more methods on real-world data, and deploying active investigations for operational validation.

Acknowledgments

ChatGPT and Claude were utilized to generate sections of this work, including text and code.

References

Hutchins, E.; Cloppert, M.; and Amin, R. 2011. Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains. Leading Issues in Information Warfare & Security Research 1.

Microsoft Threat Intelligence. 2024. Same Targets, New Playbooks: East Asia Threat Actors Employ Unique Methods. Technical Report. Redmond, WA: Microsoft Corporation.

Nimmo, B. 2015. Anatomy of an Info-War: How Russia's Propaganda Machine Works, and How to Counter It. Report. Central European Policy Institute.

Shafee, S.; Bessani, A.; Ferreira, P.M. 2024. Evaluation of LLM Chatbots for OSINT-based Cyber Threat Awareness. arXiv:2401.15127.

Terp, S.; and Breuer, P. 2022. DISARM: a Framework for Analysis of Disinformation Campaigns. In Proceedings of the 2022 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA), 1–8. Salerno, Italy. doi.org/10.1109/CogSIMA54611.2022.9830669.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; Cao, Y. 2022. ReAct: Synergizing Reasoning and Acting in Language Models. arXiv:2210.03629.