CS505 Final Project: LLM Agent for Social Media Misinformation Detection

Handi Xie

Zhichen Chen

January 22, 2024

Abstract

This project investigates an AI-Agent framework for automatic misinformation checking on social media platforms, with a focus on Reddit. The approach integrates a data pipeline for efficient content retrieval and employs Large Language Models (LLMs) for the reasoning task, specifically targeting misinformation through toxicity analysis. The findings indicate a negative correlation between toxicity scores derived from LLMs and Reddit post upvotes, suggesting that posts with higher toxicity, indicative of potential misinformation, receive fewer upvotes. This correlation underscores the potential of LLM-based methods in identifying and mitigating misinformation on social media.

Keywords: AI-Agent Framework, Misinformation Detection, Social Media, Large Language Models, Toxicity Analysis, Reddit

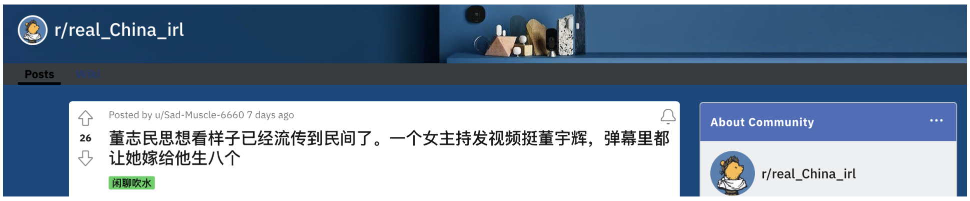
1 Team Members

Handi Xie Zhichen Chen

2 Project Description

Large Language Model powered AI-Agents are both powerful and ominous. They have been adapted for various decision-making and assistive roles. While they achieve decent performance in task-based benchmarks, their effectiveness in handling complex tasks, such as misinformation checking, is seldom tested. In this project, we propose a new Agent Framework for automatic misinformation checking, a cutting-edge development in the field of computational journalism.

Misinformation and toxic speech are proliferating on Reddit, fueled by anonymity and the platform's vast user base. These elements often manifest in misleading narratives and harmful rhetoric across various subreddits, challenging the integrity of discussions and the well-being of the community. This trend raises significant concerns about content quality and user safety. As is shown in Figure 1, it is a Reddit post in Chinese saying that "It seems Dong Zhimin's (Dong is a criminal convicted for human trafficking who have unlawfully imprisoned a woman while giving birth to eight children, and the Xuzhou eight-child mother incident has gone viral on China's internet) ideas have already spread among the people. A female host made a video supporting Dong Yuhui (A chinese influencer sharing the same Surname as Dong Zhimin), and the comments are all urging her to marry him and have eight children with him.", which is extremely offensive and toxic.



Posted by u/Sad-Muscle-6660 7 days ago

董志民思想看样子已经流传到民间了。一个女主持发视频挺董宇辉，弹幕里都让她嫁给他生八个

闲聊吹水

Figure 1: An Example of A Toxic Reddit Post

Generative AIs have been beneficial assistants in diverse scenarios, but they also contribute to the proliferation of online content, potentially impacting the health of online communities. In particular, studies have shown a surge in generated content since the launch of ChatGPT in November 2022 (Burtch et al., 2023). This project serves two purposes: firstly, it acts as an interactive web Agent framework, combining knowledge from a vector embedding database, prompt engineering, and Retrieval-augmented generation. Secondly, it is a quantitative study, exploring correlations in fake or generated news.

We use toxic speech as a surrogate for misinformation in this framework, recognizing the close relationship between the two. Toxic speech often accompanies misinformation, as it aims to manipulate or provoke emotional responses rather than present factual information. By identifying and analyzing toxic speech patterns, we can effectively target and mitigate misinformation. This approach is justified by the overlapping characteristics of misinformation and toxic speech, such as deceptive language, emotionally charged rhetoric, and the tendency to polarize or mislead audiences. As such, addressing one can significantly impact the other, making this an effective strategy in our fight against online misinformation.

3 Literature Review

Large Language Models (LLMs) like GPT-3 have revolutionized the field of artificial intelligence, demonstrating remarkable capabilities in various applications. These models have become particularly influential in decision-making and assistive roles, showcasing impressive performance across numerous task-based benchmarks. However, the effectiveness of LLMs in handling complex, real-world tasks, especially in the realm of computational journalism, remains an area of ongoing exploration.

The emergence of AI-powered agents for automatic misinformation checking represents a significant advancement in this domain. As highlighted by Augenstein et al. (2023), the current era of LLMs poses new factuality challenges. Their work, ”Factuality challenges in the era of large language models,” delves into these complexities, emphasizing the need for sophisticated approaches to tackle misinformation in the context of advanced AI technologies.

Zhou et al. (2023) further explore this theme in their study, ”Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions.” Their research, presented at the CHI Conference on Human Factors in Computing Systems, examines the nuances of AI-generated misinformation and the efficacy of both algorithmic and human-centric solutions in addressing this issue.

In the same vein, Cuartielles Saura et al. (2023) discuss the evolving role of fact-checkers in the age of ChatGPT and similar technologies. Their work, "Retraining fact-checkers: the emergence of ChatGPT in information verification," underscores the need for adapting traditional fact-checking methodologies to the challenges posed by sophisticated LLMs.

Finally, Wu and Hooi (2023) in their study, "Fake news in sheep's clothing: Robust fake news detection against LLM-empowered style attacks," address the growing concern of fake news propagation enhanced by LLMs. Their research emphasizes the development of robust detection mechanisms capable of identifying and countering sophisticated misinformation tactics facilitated by LLMs.

These studies collectively underscore the imperative of developing advanced, LLM-based frameworks for misinformation checking in the digital age, highlighting both the potential and the challenges of integrating these technologies into the field of computational journalism.

4 Data

We will be using Reddit as the data source of our project. The Reddit News subreddit, a dedicated platform for discussing current events and news stories, will serve as the primary data source for our misinformation-checking task. This subreddit is a rich repository of user-submitted news articles, discussions, and comments, offering a diverse range of topics from various global news sources. The content within this subreddit varies from mainstream media reports to user opinions and analyses, making it an ideal dataset for testing the robustness and adaptability of our Large Language Model-based misinformation-checking framework. This dynamic and varied dataset will enable us to evaluate the model's performance in real-world, complex information verification scenarios.

5 Method

In this project, we propose a new Agent Framework on automatic misinformation checking, a State-of-the-Art framework.

The two pillars of the framework are:

1. Data Pipeline: Retrieving Reddit content information storing them in proper database.

2. Reasoning: Leveraging LLM-powered AI Web Agent for summarizing and Question-Answering tasks.

See Figure 2 below for a brief System Design Flowchart.

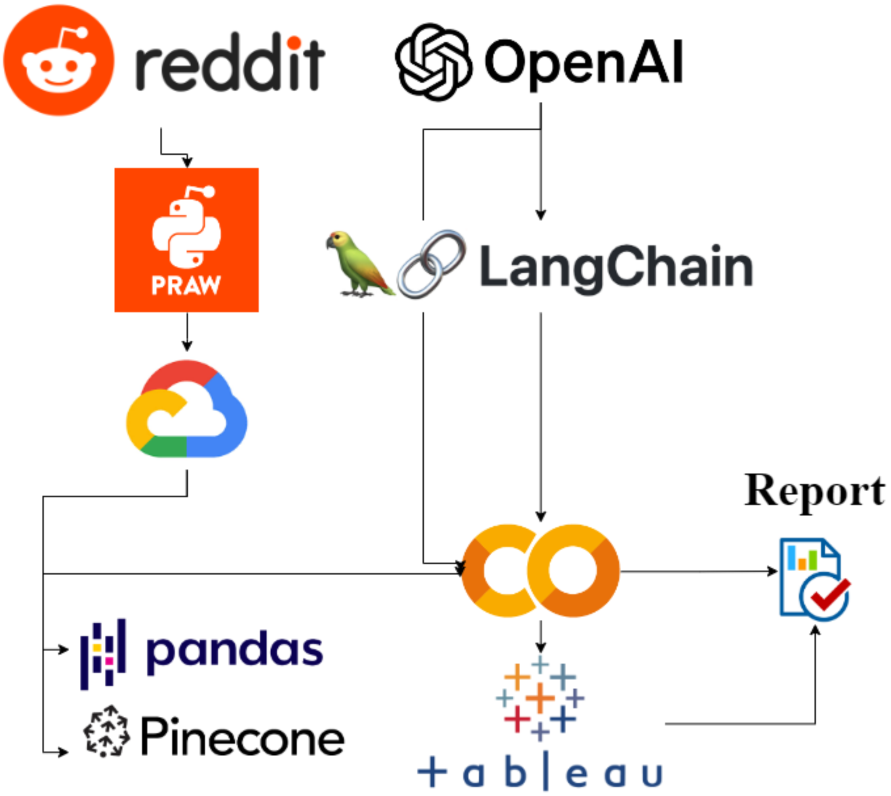


Figure 2: System Design Flowchart

5.1 Data Pipeline

The first step of our methodology involves establishing a robust data acquisition and processing pipeline. We have utilized Google Colab notebooks linked to a hosted runtime on Google Cloud to ensure scalability and efficient resource management. This setup provides the necessary computational power and flexibility required for handling large datasets and complex data processing tasks.

In light of the recent changes to the Reddit API usage terms (Reddit, 2023), our approach has been to employ PRAW (Python Reddit API Wrapper) for data retrieval. This has enabled us to systematically gather posts from the top three news and politics-related subreddits in both English and Mandarin languages, spanning the period from December 11, 2023, to December 17, 2023. The selection of these specific subreddits is strategic, as they are likely hotspots for misinformation, especially given their focus on current events and political discourse.

To comply with Reddit's API terms of service, we developed a dedicated Reddit App. This application facilitated the extraction of a wide array of data points, including post titles, timestamps, links, content, content types, upvote counts, comment counts, poster and commenter karma, commenter usernames, comment texts, and comment upvotes.

During data retrieval, we encountered several challenges unique to Reddit, such as content bans, locked threads, and limitations in accessing user information. These scenarios often resulted in "Exceptions: Query not found" errors. To address this, we implemented robust exception handling procedures that allowed the system to gracefully manage API limits and timeouts, ensuring continuous data collection without significant interruptions.

The data storage process involved meticulous data cleaning using the pandas library. This step was crucial for filtering out extreme cases and anomalies that could skew the analysis. After cleaning, we embedded the post contents into a vector database using PineCone. This not only facilitated efficient storage but also laid the groundwork for advanced analytical techniques. PineCone's vector database was specifically chosen for its ability to handle high-dimensional data and its optimization for similarity search, which are essential features for our subsequent misinformation detection algorithms.

By leveraging these technologies and methodologies, we have established a comprehensive and robust framework for collecting and processing social media data. This framework is fundamental to our overall objective of detecting and analyzing misinformation on social media platforms using Large Language Model (LLM) agents.

5.2 Reasoning Task for Misinformation Detection

For reasoning tasks (Misinformation Detection), with LangChain and OpenAI APIs we have built from scratch a LLM-powered AI Agent for webpage summarization and toxicity detection tasks. For optimization purposes, we have used chunking techniques to summarize webpage tokenized context frame larger than LLM model's token frame; used Retrieval Augmented Generation (RAG) to set sentiment for better performance in few-shot learnings.

The development began with the creation of an application on the OpenAI platform, a necessary step to access OpenAI's powerful language models. Obtaining an API key was crucial as it enabled the interaction with GPT-3.5 and GPT-4 models, known for their advanced capabilities in language understanding and generation.

Building the AI agent was a complex process, involving the use of LangChain and OpenAI APIs. The AI agent was equipped with the UnstructuredURLLoader, allowing it to process content directly from web pages. This was a significant feature as it enabled the agent to handle real-world data effectively.

To ensure the AI agent worked efficiently, significant emphasis was placed on preprocessing the content. This included optimizing the context length by trimming unnecessary whitespace and punctuation, a step that ensured consistency and accuracy in the model's outputs.

Handling lengthy web articles presented a unique challenge due to the token limits of LLM models. To address this, we employed a chunking technique. This involved breaking down longer articles into smaller chunks, summarizing each chunk individually, and then combining these summaries into a final comprehensive overview. This approach was not only effective in managing content length but also in ensuring that no crucial information was lost in the summarization process.

The setting of the temperature parameter to zero was a strategic choice, aimed at reducing randomness and enhancing the determinism of the outputs. This was particularly important for maintaining the reliability of the information, especially in tasks involving misinformation detection.

The choice of GPT-3.5 and GPT-4 models was deliberate, considering their status as the most advanced in the field. These models have been refined through reinforcement learning from human feedback (RLHF) by OpenAI, which has significantly improved their performance and relevance to real-world tasks. We also undertook a comparative analysis of different models to determine the most effective setup for our specific needs. An example of our prompt for toxic speech detection is as follows.

> Please evaluate the following content for its level of toxicity. Provide a rating on a scale from 0 to 10, where 0 indicates no toxicity (completely neutral or positive content), and 10 indicates extremely toxic content, such as content that is harmful, offensive, or incites negativity. Base your assessment on the tone, language, and potential impact of the content. Do not write anything but a 0-10 score.

Lastly, the integration of OpenAI's ChatCompletion feature into our AI agent was crucial. This feature excels in handling conversational contexts and generating coherent, contextually appropriate responses. It was instrumental in enabling our AI agent to interact with and analyze social media content effectively, particularly in identifying and assessing misinformation.

Through the integration of these advanced techniques and tools, our AI agent is not just capable of efficient webpage summarization but is also adept at detecting toxic content, which often accompanies misinformation. This integral approach is pivotal in our efforts to combat misinformation on social media platforms using the latest advancements in AI technology.

6 Analysis and Results

In our study, we employed Tableau and Seaborn for sophisticated data visualization and presentation purposes. As depicted in Figure 3, our analysis reveals a negative correlation between the Large Language Model (LLM)-based toxicity scores and the percentage of post upvotes. This inverse relationship suggests that employing LLM for misinformation detection achieves satisfactory accuracy. The toxicity scores, calculated using LLM, have demonstrably succeeded in accurately reflecting the toxicity levels in subreddit posts. Utilizing metrics such as Post Upvote Percentage (which measures the distribution of upvotes among posts with varying toxicity levels within a given subreddit) and Comment Percentage (which assesses the distribution of replies to posts across different toxicity levels), we observed a notable trend. Subreddits characterized as echo chambers, such as 'real_China_irl', showed a higher propensity for toxic posts. This pattern underscores how these echo chambers contribute to and perpetuate a toxic environment.

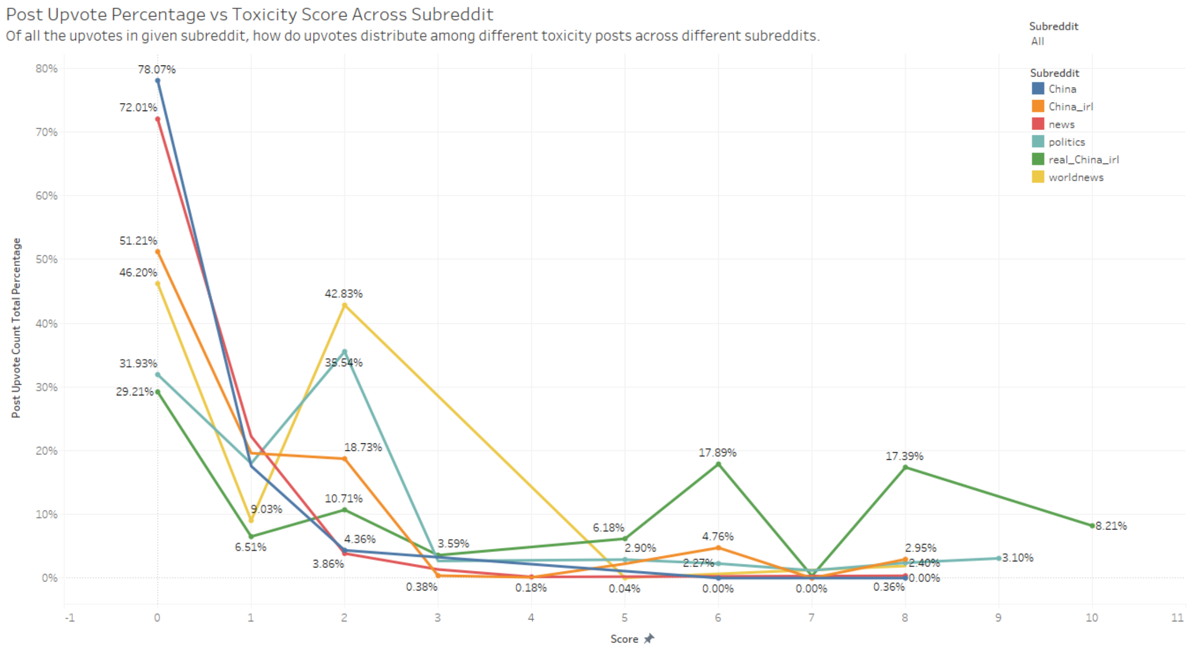


Figure 3: Correlation between Toxicity Score and % of post upvote

7 Implication and Contribution

7.1 Implication

The challenge of detecting misinformation is a significant hurdle for both content platform moderators and users. Misinformation, often characterized by its deceptive and manipulative nature, poses a threat to the integrity of information dissemination and can have far-reaching consequences, including the shaping of public opinion and influencing political and social discourse. Its elusive nature, coupled with the sophisticated tactics used by spreaders of misinformation, makes it particularly difficult to identify and counter.

In our approach, we leverage the state-of-the-art capabilities of GPT-4, an advanced iteration of OpenAI's generative language models, for the detection of misinformation. The rationale behind employing GPT-4 lies in its sophisticated language processing abilities, which include understanding context, detecting nuances in text, and discerning patterns indicative of misinformation. GPT-4's extensive training on diverse datasets enables it to recognize a broad range of misinformation strategies.

Our results demonstrate that the constructed toxicity score, derived from GPT-4's analyses, correlates with the number of post upvotes on Reddit. This correlation suggests that posts with higher toxicity scores, indicative of potential misinformation, tend to receive fewer upvotes, reflecting user engagement and perception. Consequently, this finding supports the effectiveness of using an LLM-based approach as a desirable method for misinformation detection.

7.2 Contribution

Our research contributes significantly to the field of misinformation detection. We demonstrate the practical application of GPT (Generative Pre-trained Transformer) models, specifically GPT-4, in identifying misinformation across social media platforms. This method stands out not only for its effectiveness in English but also in its adaptability to multiple languages, such as Chinese. The ability to process and analyze content in various languages is crucial, considering the global nature of misinformation spread and the diverse linguistic backgrounds of social media users.

By showcasing the effectiveness of GPT-4 in detecting misinformation in different languages, our work opens avenues for more inclusive and comprehensive monitoring of online content. This multilingual capability is particularly important in the context of global platforms like Reddit, where misinformation can transcend linguistic barriers and affect a wide audience. Our method thus provides a scalable and versatile tool for content moderators and platforms in their ongoing efforts to combat the spread of misinformation and maintain the integrity of online discourse.

8 Statement of Contribution

In this research project, two team members (Handi Xie and Zhichen Chen) have collaborated closely, contributing jointly to each pivotal phase. During the ideation stage, both members engaged in brainstorming and conceptualizing the project's framework, setting a solid foundation for the research. The literature review was a collaborative effort, with both members diligently examining existing studies and theories to inform our approach.

Subsequent stages, including data retrieval, coding, and results analysis, witnessed a seamless integration of both members' skills and insights. Each member played a vital role in gathering data, implementing coding strategies, and dissecting the results to draw meaningful conclusions. The final stage of report writing was a culmination of this joint effort, where both members amalgamated their findings and perspectives into a comprehensive and insightful report, reflecting their shared commitment and contributions to the project.

References

Augenstein, I., T. Baldwin, M. Cha, T. Chakraborty, G.L. Ciampaglia, D. Corney, R. DiResta, E. Ferrara, S. Hale, A. Halevy, et al. 2023. Factuality challenges in the era of large language models. *arXiv preprint arXiv:2310.05189* .

Burtch, G., D. Lee, and Z. Chen. 2023. The consequences of generative ai for ugc and online community engagement. *Available at SSRN 4521754* .

Cuartielles Saura, R., X. Ramon Vegas, and C. Pont Sorribes. 2023. Retraining fact-checkers: the emergence of chatgpt in information verification .

Reddit. 2023. Reddit data api terms. `https://www.redditinc.com/policies/data-api-terms`. Accessed: 2023-12-22.

Wu, J. and B. Hooi. 2023. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. *arXiv preprint arXiv:2310.10830* .

Zhou, J., Y. Zhang, Q. Luo, A.G. Parker, and M. De Choudhury 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–20.