Misinformation Identification: An Agentic RAG Based Technique

Dissertation Stage-I

M.Tech
in
Computer Science and Engineering

by

Jimmy Aghera
(23535031)

Under the Guidance of
Prof. Durga Toshniwal

Department of Computer Science and Engineering
Indian Institute of Technology, Roorkee
May 2024

Candidate Declaration

I hereby certify that the work, which is being presented in the project report, entitled Misinformation Identification: An Agentic RAG Based Technique, in partial fulfillment of the requirement for the award of the Degree of Master of Technology in Computer Science & Engineering and submitted to the institution is an authentic record of my own work carried out under the supervision of Professor Durga Toshniwal.

Date: ____________________

Jimmy Aghera
(23535010)

Prof. Durga Toshniwal

Abstract

This report on Misinformation Identification: An Agentic RAG Based Technique highlights the importance of addressing the spread of inaccurate information on social media platforms. As digital platforms continue to shape opinions and decisions, ensuring the reliability of shared content has become a significant area of interest.

In this report, we explore advanced methodologies for misinformation detection, focusing on the application of Retrieval-Augmented Generation (RAG) and large language models. By integrating retrieval-based information sourcing with generative AI, we aim to build a framework that verifies social media claims effectively. This approach considers factors such as claim context, factuality, and source reliability to deliver accurate verification. Through this solution, we strive to contribute to a more informed and trustworthy digital environment.

Table of Contents

List of Figures

1 Introduction

In the digital age, the rapid dissemination of information across social media and online platforms has transformed how people consume news and share opinions. However, this ease of communication has also led to an alarming rise in misinformation, which can have severe societal, political, and economic repercussions. From fabricated news articles to manipulated media, misinformation spreads faster than factual content[8], often influencing public opinion and decision-making.

To address this pressing challenge, the proposed approach leverages Retrieval Augmented Generation (RAG), a state-of-the-art technique combining Large Language Models (LLMs) with external knowledge retrieval systems. By incorporating agent capabilities, this method retrieves relevant and verified information to evaluate the correctness of a claim in real-time. This technique enables the identification of misleading content with higher accuracy and contextual understanding compared to traditional methods.

This report focuses on designing and implementing a misinformation detection system that analyzes social media content, particularly tweets and articles, to determine their authenticity. By utilizing advanced natural language processing(NLP) techniques, the system processes claims, retrieves supporting evidence from trusted sources, and evaluates the claim's validity. The aim is to curb the spread of misinformation by providing users with reliable insights and empowering platforms to make informed decisions.

Through this research, we seek to bridge the gap between automation and accountability in combating misinformation, paving the way for a safer and more reliable digital information ecosystem.

1.1 Motivation

The motivation behind addressing misinformation in India arises from its significant societal and democratic impact. A 2024 study revealed that 65% of first-time voters in India encountered fake news, particularly via platforms like WhatsApp and Instagram, influencing political perceptions and choices. This highlights the urgent need for media literacy and effective mechanisms to combat misinformation, especially with India being one of the most affected nations globally due to its vast and diverse online user base.

India's 2024 elections demonstrated how misinformation, including deepfakes and fake news about election processes, spread across regional languages and multiple platforms, eroding trust in democratic institutions. Efforts like collaborative fact-checking initiatives have shown promise, but the complexity of the issue necessitates advanced technological solutions.

1.2 Problem Statement

Traditional misinformation detection methods, relying on static databases or manual fact-checking, are unable to cope with the speed, scale, and evolving nature of false narratives.

Additionally, the absence of context-aware and dynamic systems for misinformation identification leads to challenges in addressing diverse and complex information ecosystems. This inadequacy results in delayed detection, widespread influence of false content, and erosion of trust among individuals and institutions.

To address this issue, there is a pressing need for an advanced system that can analyze, retrieve, and verify information in real time. Leveraging Retrieval-Augmented Generation (RAG) with agentic capabilities offers a promising solution to dynamically identify and counter misinformation while providing evidence-based insights. This study aims to design and implement such a system to mitigate the growing impact of misinformation on society.

1.3 Organization of report

In this report for combating misinformation, the section 2 have Multiple literature review which will reveal some techniques which are being proposed for this problem statement, after that in Section 3 will discuss about proposed methodology, after that section 4 and 5 will discuss about current Experiments and conclusions.

2 Literature Review

2.1 Methodologies

Misinformation detection initially relied on traditional neural network techniques, such as Hybrid Convolutional and Bi-LSTM models, which effectively processed textual data in controlled environments. Surendran et al. (2021)[6] proposed a COVID-19 misinformation detector using these methods, highlighting their ability to identify fake news with reasonable accuracy. However, these models lacked the capacity to understand long-range dependencies in text, limiting their scalability and effectiveness in complex scenarios.

The advent of transformer architectures significantly enhanced the field by enabling models to capture contextual relationships in large text datasets. Retrieval-Augmented Generation (RAG) techniques, which combine retrieval mechanisms with generative capabilities, addressed key limitations of earlier models. Hang et al. (2024)[3] introduced TrumorGPT, a RAG-based framework leveraging Wikipedia for claim validation, emphasizing its robustness in handling diverse misinformation. Similarly, Yue et al. (2024)[7] focused on integrating evidence-driven retrieval for online misinformation, highlighting the importance of domain-specific knowledge sources for accurate detection.

Recent advancements include leveraging LLMs for misinformation detection through augmented claim-matching mechanisms. Choi and Ferrara (2024) developed Fact-GPT[1], which matches claims against pre-verified debunked data, demonstrating the utility of fine-tuned LLMs in public health misinformation contexts, such as COVID-19. They emphasized the potential of smaller LLMs, such as Llama-2, for cost-effective and scalable solutions. Works like those by Singal et al. (2024)[5] integrated RAG with few-shot learning, improving evidence-backed fact-checking efficiency. Similarly, Ram et al. (2024)[4] presented CrediRAG, a Reddit-specific framework using credibility-based retrieval, which demonstrated success in network-augmented misinformation detection.

2.2 Research Gap

Traditional machine learning models[6] and recurrent neural networks face significant challenges in processing the complexity of human language. These models often fail to understand the contextual nuances within text, particularly for longer and more intricate sentences, leading to suboptimal performance in misinformation detection tasks.

While fine-tuning pre-trained models on domain-specific datasets has proven effective, such approaches are limited by the quality and size of the training datasets. These datasets are often outdated, reducing the accuracy and reliability of the models over time, as evidenced by several studies in the literature[1].

RAG based systems have shown promise in handling misinformation. However, existing implementations like TrumorGPT[3] lack robust evidence-based fact-checking mechanisms. This undermines their trustworthiness, as users require transparent and reliable validation of the generated outputs.

Most RAG-based systems[5] rely on top-K retrieval methods to fetch relevant information. While efficient, this approach may compromise the reliability and comprehensiveness of the retrieved data, potentially leading to incomplete or misleading conclusions.

3 Proposed Methodology

3.1 Methodology

The proposed methodology seeks to address the research gap outlined in Section 2.2 by going beyond the use of RAG techniques alone to counter misinformation. It combines RAG with agentic capabilities of Large Language Models (LLMs). RAG will be used to identify and address claims that have already been debunked by well-known open-source fact-checking organizations. If a user's claim does not match any existing fact-checker data, the agent and its tools will search the internet to verify whether the claim is accurate or falsely presented. This approach ensures a more comprehensive and dynamic solution for misinformation detection as shown figure 1.

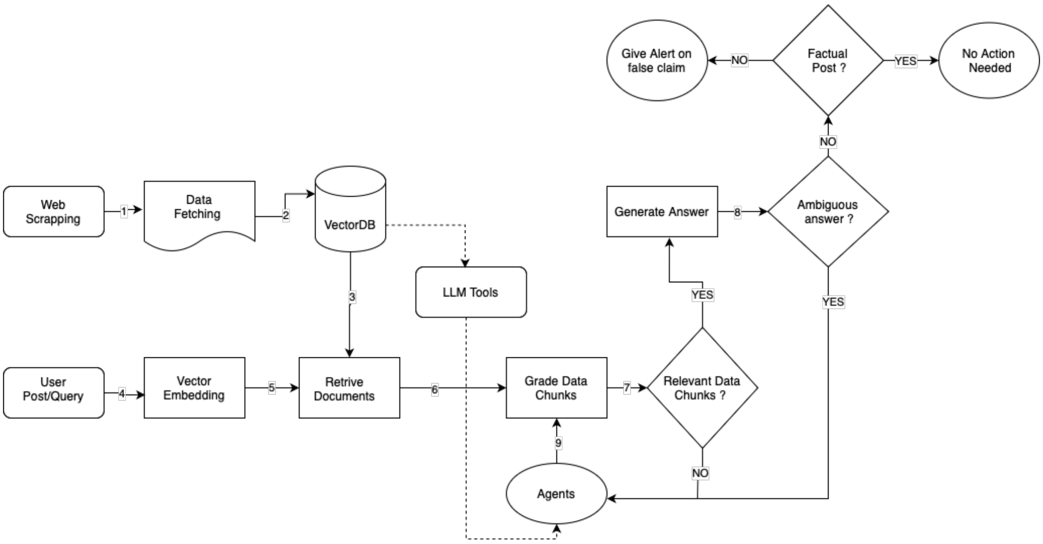

Figure 1: Architecture of proposed Method to counter misinformation.

3.1.1 RAG module

The Retrieval-Augmented Generation (RAG) module serves as the backbone for factual information retrieval in this methodology. It enhances the LLM's ability to access relevant data by utilizing a vector database, such as Chroma, instead of traditional knowledge graphs. The vector database enables more efficient retrieval using techniques like Locality Sensitive Hashing (LSH), allowing faster access to data for evaluating misinformation. For claims that have already been debunked by trusted fact-checking websites such as Logically, Boom, Ptinews, and Altnews, the RAG system fetches information directly from these sources. When a user submits a query (e.g., a social media post or claim), the RAG system performs a similarity search on the vector database to retrieve relevant data chunks. These retrieved chunks are then injected into a pre-trained language model like GPT-4, which evaluates the factual accuracy of the query. If no matching data is found in the database, the system triggers the Agent module for further investigation.

3.1.2 Agent module

The Agent module supplements the RAG system by leveraging LLM agents to perform active fact-checking when the RAG module cannot find sufficient information in the vector database. This module uses tools to query external sources in real-time, such as websites, news outlets, and fact-checking databases, to verify the claim's accuracy. The LLM agent is capable of autonomously searching across multiple platforms to check the claim's validity. In cases where contradictory or ambiguous information is retrieved, the agent further iterates on the query, searching for more conclusive evidence until a reliable answer is found. The agent also utilizes the vector database itself as one of the search tools, ensuring faster retrieval during these checks. The inclusion of source links in the metadata further supports the transparency and reliability of the fact-checking process, making the results more credible for users. This agentic structure ensures that even when pre-existing knowledge in the vector database does not suffice, the system can still actively work towards a definitive conclusion.

3.2 Dataset

The dataset utilized in this study is sourced primarily from logicallyfacts, a fact-checking website. This dataset is raw data being scraped for the use of the vector database. The required columns are Context, Verdict, and Link, which are common factors across all websites that will be scraped in the future. This uniformity ensures that the RAG system can seamlessly adapt to any source, enabling efficient identification of misinformation.

The dataset currently contains **8,475 entries** and includes the following attributes:

- **Topic (string):** Provides a brief overview or title describing the central claim or discussion point.
- **Author (string):** Person who fact-checked.
- **Date (date):** Specifies the date when the claim was fact-checked.

- **Category (enum):** Categorizes claims into one of the following topics:
  - `climate, conspiracies, economics, events, geopolitical, health, human-rights, media, politics, tech, conflict, sports`.
- **Context (string):** Provides detailed background information or supporting content for the claim.
- **Verdict_Status (enum):** Indicates whether the claim has been verified, with possible values:
  - `Fact-Check, Unknown`.
- **Verdict (enum):** Represents the judgment of the claim, with possible values:
  - `False, Misleading, Partly_True, True, Fake`.
- **Link (url):** URL of the page which is being accessed for the given claim.

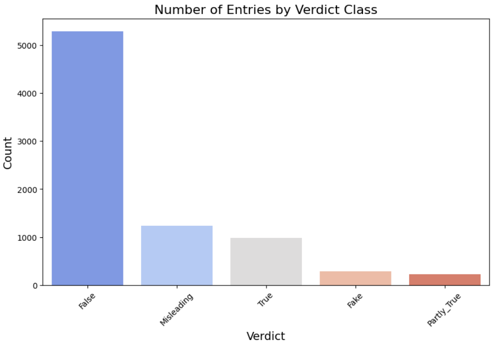

Figure 2: Number of data for each verdict class

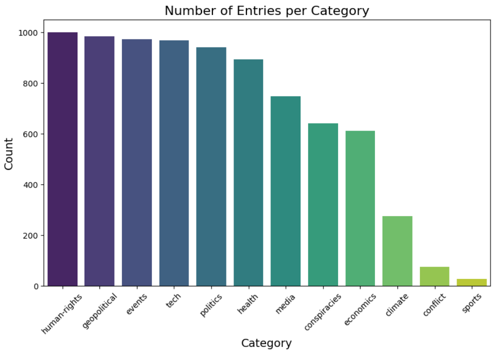

Figure 3: Number of data for each Category class

This dataset serves as the foundation for implementing the RAG framework, allowing it to retrieve information based on similarity searches and provide users with accurate, contextually relevant fact-checking results. Future deployments aim to expand the dataset by incorporating additional entries from other credible fact-checking sources.

4 Experiments

The current experiment focuses on evaluating the accuracy of the standalone RAG system and identifying the potential need for integrating agents on top of the RAG framework. Preliminary results reveal that while the RAG system performs effectively in retrieving factual claims, it is prone to Type-II errors, particularly when handling ambiguous or nuanced misinformation.

This issue is not unique to this implementation but has been widely observed across similar approaches. For instance, existing methods such as those proposed in [3], [7], and [2] also exhibit challenges in minimizing Type-II errors, emphasizing the importance of incorporating additional layers of intelligence, such as agents, to enhance decision-making and fact-checking accuracy.

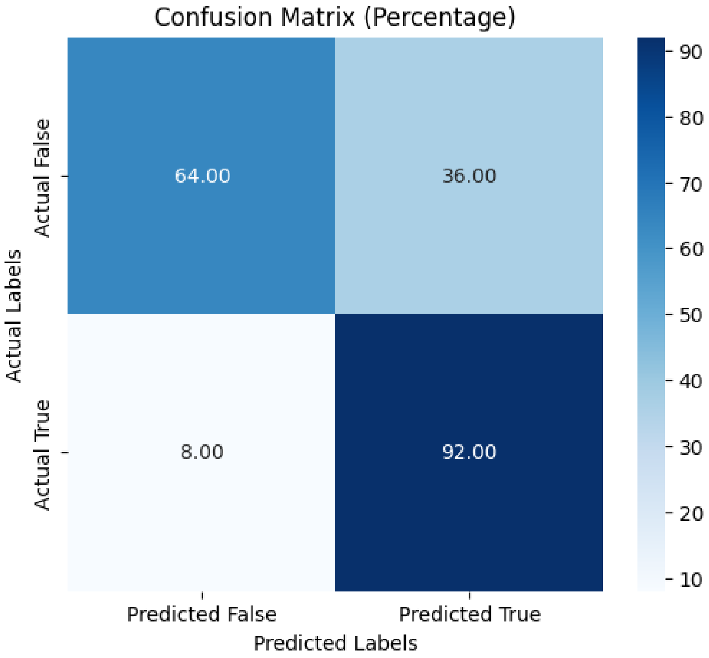

Figure 4: Confusion matrix for only RAG based method

An analysis of the data reveals that a significant contributor to the occurrence of Type-II errors is the use of techniques such as ”astroturfing” or ”disinformation sandwiching.” These strategies involve hiding misinformation within a mix of factual information, making it difficult for standalone RAG systems to accurately differentiate between true and false claims.

This challenge, as highlighted in the observed data, underscores the limitations of existing methods. The integration of agents, as detailed in the proposed methodology(3), offers a potential solution. Agents can leverage advanced reasoning capabilities to scrutinize claims more thoroughly, effectively identifying and addressing misinformation even when it is embedded within layers of factual content.

> Nine different subsidies that the U.S. government gives to an industry that makes more money than any other industry, including refunds for drilling costs and refunds to cover the cost of searching for oil. Subsidies for oil and gas companies make up 88 percent of all federal subsidies. Just cutting the oil and gas subsidies out would save the U.S. government $45 billion every year.

In this example, all the statements are correct except the last one. The phrase *every year* is incorrect; the correct term should be *ten years*. This discrepancy significantly alters the perception of the claim and its impact.

One of the key advantages of the Agentic RAG system is its ability to decompose complex claims into smaller sub-claims and verify them iteratively through multiple queries to the database or the internet. This ensures a higher level of precision in identifying misinformation by cross-checking every micro-claim.

Consider the input claim:

> *"The average global temperature in 2023 was the highest ever recorded."*

Decomposition by Agent:

- **Sub-claim 1:** *"What was the average temperature in 2023?"*
- **Sub-claim 2:** *"Was this the highest ever recorded?"*

Search Queries and Results:

- **Result 1:** *"1.2°C higher than pre-industrial levels in 2023."*
- **Result 2:** *"2016 and 2019 were the hottest years before 2023."*

Reasoning by Agent:

- Sub-claim 1 resolved as **True**: *"2023 temperature: 1.2°C higher."*
- Sub-claim 2 resolved as **True**: *"Higher than 2016 and 2019."*

Final Output:

> *"Combined analysis confirms the claim as* ***True****."*

5 Discussion

5.1 Future Work

Comparative experiments will be conducted to evaluate the time efficiency and accuracy of the RAG approach versus the Agentic RAG approach. This will provide quantitative insights into the benefits of incorporating LLM agents into the misinformation detection process.

The integration of multilingual and image-processing agents will enable the system to handle claims presented in diverse languages and visual formats. This will expand the applicability of the system to address misinformation in global contexts, including claims shared as memes, infographics, or non-textual content.

The performance will be measured against baseline algorithms, focusing on accuracy improvements. These evaluations will demonstrate the efficacy of the proposed methodology in identifying and debunking misinformation with greater precision.

5.2 Conclusion

This study introduces an innovative methodology for misinformation detection by combining the RAG framework with LLM-based agents. The approach leverages fact-checking datasets and advanced query mechanisms to effectively address misinformation.

The RAG module efficiently retrieves verified claims from a vector database, while the agent module handles ambiguous queries or claims not covered by existing datasets. This dual mechanism ensures robust adaptability and enhanced fact-checking capabilities.

This method demonstrates significant potential for combating misinformation, contributing to more reliable information dissemination across digital platforms.

References

[1] Eun Cheol Choi and Emilio Ferrara. Fact-gpt: Fact-checking augmentation via claim matching with llms. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 883–886, 2024.

[2] Francesco Bombassei De Bona. Integrating retrieval augmented generation in large language models for effective fact-checking. 2024.

[3] Ching Nam Hang, Pei-Duo Yu, and Chee Wei Tan. Trumorgpt: Query optimization and semantic reasoning over networks for automated fact-checking. In *2024 58th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2024.

[4] Ashwin Ram, Yigit Ege Bayiz, Arash Amini, Mustafa Munir, and Radu Marculescu. Credirag: Network-augmented credibility-based retrieval for misinformation detection in reddit. *arXiv preprint arXiv:2410.12061*, 2024.

[5] Ronit Singal, Pransh Patwa, Parth Patwa, Aman Chadha, and Amitava Das. Evidence-backed fact checking using rag and few-shot in-context learning with llms. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 91–98, 2024.

[6] Pranav Surendran, B Navyasree, Harshitha Kambham, and M Anand Kumar. Covid-19 fake news detector using hybrid convolutional and bi-lstm model. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 01–06. IEEE, 2021.

[7] Zhenrui Yue, Huimin Zeng, Yimeng Lu, Lanyu Shang, Yang Zhang, and Dong Wang. Evidence-driven retrieval augmented response generation for online misinformation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5628–5643, 2024.

[8] Yizhou Zhang, Karishma Sharma, Lun Du, and Yan Liu. Toward mitigating misinformation and social media manipulation in llm era. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1302–1305, 2024.