A Large Language Model-based Fake News Detection Framework with RAG Fact-Checking

Yangxiao Bai and Kaiqun Fu
Department of Computer Science
South Dakota State University
Brookings SD, U.S.A.
{bai.yangxiao,kaiqun.fu}@sdstate.edu

2024 IEEE International Conference on Big Data (BigData) | 979-8-3503-6248-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/BigData62323.2024.10826000

***Abstract*—The widespread dissemination of online misinformation poses significant threats to the public interest, highlighting the urgent need for effective fake news detection. In the era of Large Language Models (*LLM*s), the rise of *AI*-generated fake news has intensified this issue, making misinformation more pervasive and harder to control. While fact-checking offers a promising solution by leveraging external knowledge, efficiently linking claims within news articles to relevant external facts remains a significant challenge. To address this, we propose a misinformation detection framework *FCRV* (Full-Context Retrieval and Verification) that constructs a "full-context" for news articles by integrating *LLM*-based claim extraction with Retrieval-Augmented Generation (*RAG*) for fact-checking. We implemented an *LLM* pipeline for human-like extraction of key claims from datasets, significantly improving extraction quality over traditional methods. Our retrieval workflow effectively detects fictitious entities prevalent in *AI*-generated news by identifying claims lacking a basis in reality. Experiments across multiple human-generated and *AI*-generated datasets demonstrate that verifying news using this "full-context" approach leads to more stable and robust fake news detection, enhancing scalability, accuracy, and the model's ability to handle *AI*-generated content.**

I. Introduction

The rapid and widespread dissemination of online misinformation presents a significant threat to society, impacting public opinion and behavior on a global scale. The persistent and evolving nature of fake news has intensified the demand for reliable and efficient detection methods capable of addressing this pressing issue. Compounding these challenges, the emergence of generative *AI* has introduced a new dimension to the misinformation landscape. Sophisticated models like *GPT*-3 generate content that closely mimics human writing, making it difficult to distinguish between machine-generated and human-authored text and thereby challenging traditional detection methods such as writing-style recognition [6]. As a result, there is a pressing need for detection strategies that can contend with the complexities introduced by *AI*-generated misinformation.

Fact-checking has emerged as a highly effective approach for early detection of fake news. By cross-referencing claims against reliable sources, fact-checking provides a robust means of combating misinformation. However, existing methods face challenges when dealing with *AI*-generated fake news. Some approaches rely solely on internal textual features without leveraging external knowledge, limiting their effectiveness against sophisticated misinformation [2]. Others enhance language models with factual knowledge by pre-training on specific corpora [3], but may not capture the dynamic nature of real-world information.

Building on these insights, we propose a full-context approach that expands upon established fact-checking frameworks to better address the complexities of *AI*-generated fake news. Our approach involves three key components: (1) extracting claims from the article, (2) retrieving relevant external information that substantiates or refutes each claim, and (3) modeling the relationships between these claims and external sources. This comprehensive framework is particularly effective against *AI*-generated fake news, which often fabricates events or entities with no basis in reality. The "absence of evidence" signal, captured through our retrieval process, becomes a crucial marker for detecting such misinformation. To the best of our knowledge, we are the first to leverage the non-existence of corroborative data as proof of falsification.

To effectively construct this full-context, our approach integrates advanced techniques—specifically *LLM*s and Retrieval-Augmented Generation (*RAG*). *LLM*s are more flexible and adaptable than traditional NER tools like *TagMe* [1], which rely on predefined entity databases, making LLMs more effective in dynamic or unstructured content. *RAG* enhances fact-checking by aligning claims with external sources, making the detection process more comprehensive and efficient. Our contributions are threefold:

1) We implement an *LLM* pipeline that enables human-like extraction of claims from datasets, significantly enhancing the quality of extraction compared to traditional methods. This advancement contributes to constructing the "full-context" of news articles.
2) We develop a retrieval workflow that effectively tackles the prevalent issue of fictitious entities in *AI*-generated news, enhancing the detection of content that lacks a basis in reality.
3) We conduct experiments across multiple human-generated and *AI*-generated datasets, demonstrating that *LLM*s performing verification within the "full-context" framework are more stable and robust. We also make our extraction results publicly available to support further research.

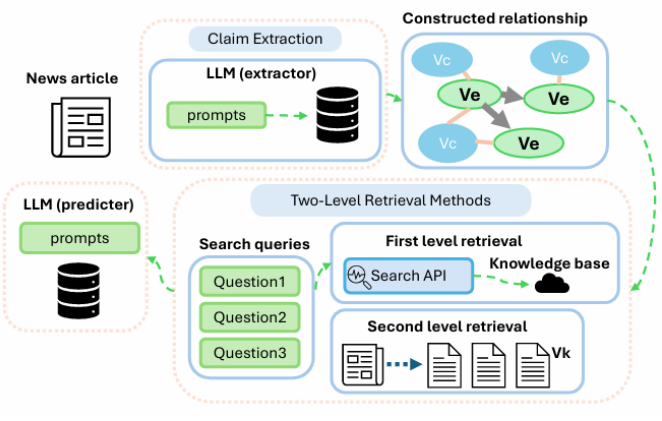

Fig. 1. overall structure of the framework

II. Problem Statement

Each news article in our dataset $A = \{A_1, A_2, \ldots, A_N\}$ contains text from which we extract triple relationships in the form of (`subject-predicate-object`) or (`subject-attribute-value`) using a Large Language Model (*LLM*). Our goal is to train an *LLM* that leverages these extracted relationships, along with external knowledge retrieved via online search using Retrieval-Augmented Generation (*RAG*), to accurately classify news articles as fake ($y = 0$) or true ($y = 1$). For each article A_i, the *LLM* extracts key claims and entities to form a set of triples T_i, and we perform online searches to obtain pertinent external knowledge K_i and record retrieval behaviors such as entities not found or contradictory information, denoted as B_i. The *LLM* processes the original article A_i, the extracted triples T_i, the retrieved external knowledge K_i, and the retrieval behaviors B_i to perform fact-checking and reasoning. The primary objective is to learn a function f that maps this enriched input to the corresponding label y, *i.e.*, $f : (A_i, \mathrm{T}_i, \mathrm{K}_i, \mathrm{B}_i) \rightarrow y$, where $y \in \{0, 1\}$. Furthermore, we retrain the *LLM*'s reasoning abilities on a dataset containing *AI*-generated news to enhance its capability to handle synthetic content. By integrating *LLM*s with *RAG* and incorporating retrieval behaviors, our method aims to improve the model's ability to capture key information in fake news, reduce training time, and enhance classification accuracy.

III. Methodology

Our misinformation detection framework *FCRV* (Full-Context Retrieval and Verification) comprises three key modules: Claim Extraction and Processing, Integrated Retrieval Approach, and Reasoning-Enhanced Verification.

A. Claim Extraction and Processing

Accurate identification of claims and entities within news articles is crucial for effective misinformation detection. Traditional text-matching tools often struggle with issues like abbreviations, misspellings, and the subtleties of natural language, which can hinder precise entity recognition and linking. With advancements in *LLM*s and their enhanced semantic understanding, we can now perform more accurate and human-like annotations. In this stage, We design task-specific prompts that enable the *LLM* to identify and extract T_i, along with their contextual information within the articles. This extraction process lays the foundation for subsequent analysis by providing a structured representation of the article's content, which is essential for effective misinformation detection.

B. Integrated Retrieval Approach

Our retrieval process operates in two levels to ensure comprehensive fact-checking. In the first level, we perform online searches using the DuckDuckGo Search API[1], provided by the LangChain[2] community. For each extracted claim T_i, we formulate specific queries and retrieve the top five search results, storing all retrieved data in a database. This step gathers a broad range of potentially relevant information to construct $K_i$for subsequent analysis. To align the search with our dataset's context, we restrict searches to English in the United States (us-en) and apply filtering rules to exclude content from certain websites[3] that may contribute to our dataset. In the second level, the *LLM* matches the extracted claims and entities with the most semantically similar chunks from the embedded documents. During this stage, we log retrieval behaviors for each claim and entity to capture signals B_i indicative of potential misinformation. These behaviors include successful retrieval when substantial supporting information is found; no information found when the *LLM* cannot find relevant information in the embedded documents; contradictory information when the retrieved content directly contradicts the claim or entity; low information density when only minimal information is available, indicating obscurity; and source credibility issues when the retrieved information comes from sources lacking credibility or authority.

C. Reasoning-Enhanced verifying

In the final stage, we perform reasoning-enhanced retraining to improve the model's ability to assess the veracity of news articles. Using the enriched input which includes the original article, extracted claims and entities, retrieved external knowledge, and logged retrieval behaviors—the *LLM* conducts fact-checking and reasoning. The *LLM* integrates contextual information from the article with external evidence and the associated retrieval behaviors for each claim and entity. It considers factors such as contradictory information, absence of evidence, low information density, and source credibility issues to determine the truthfulness of the content. This process enables the *LLM* to classify articles as true or fake and provide explanations based on the retrieval behaviors.

IV. Experiment

A. Dataset

Our dataset is derived from *GossipCop++* and *PolitiFact++*, introduced by Su *et al.* [5], containing human-written

[1]https://duckduckgo.com/api
[2]https://www.langchain.com/
[3]Excluded websites include `politifact.com`, `gossipcop.com`, `snopes.com`, `factcheck.org`, and `suggest.com`

TABLE I
MODEL PERFORMANCE ON GOSSIPCOP AND OUT-OF-DOMAIN SETS (AI-GENERATED AND POLITIFACT)

Model	Full-Context	GossipCop		AI-Generated		PolitiFact	
		Accu	F1	Accu	F1	Accu	F1
H-LSTM	No	78.3%	77.5%	68.8%	70.2%	50.0%	52.0%
BERT	Yes	82.9%	83.5%	83.0%	82.8%	60.5%	61.0%
RoBERTa	Yes	84.5%	85.0%	84.0%	84.2%	62.0%	63.0%
ALBERT	Yes	83.2%	83.8%	83.5%	83.0%	61.5%	62.0%
DeBERTa	Yes	85.0%	85.5%	84.8%	85.0%	63.0%	64.0%
FCRV	Yes	88.0%	89.4%	86.4%	88.5%	71.6%	82.8%

fake (*HF*) and real news (*HR*) articles from *FakeNewsNet* [4], filtered to include articles with both a title and description. For out-of-domain testing, we use two datasets: (1) the *AI-generated* dataset, which includes machine-paraphrased real news (*MR*) and machine-generated fake news (*MF*) produced by Su *et al.* [5] using *ChatGPT* and Structured Mimicry Prompting; (2) the human-written dataset, comprising 97 real and 97 fake news articles sampled from *PolitiFact++*. All experiments were conducted using NVIDIA A100 GPUs.

B. Evaluation Metrics

In this balanced training and testing datasets, subclass-wise accuracy serves as the primary evaluation metric, providing a focused and informative assessment of model performance within each specific subclass. This measure allows us to explore potential internal biases in the detector, as well as its accuracy in correctly classifying both fake and real news. Due to the balanced nature of our dataset, additional metrics such as *F1 score*, *precision*, *recall*, and overall *accuracy* can be easily derived from subclass-wise accuracy.

C. Baseline

To evaluate the effectiveness of our proposed method, we compare it against several baseline models commonly used in fake news detection and text classification tasks, adjusted to incorporate the full-context approach where applicable. These baselines include Hierarchical LSTM (*H-LSTM*), which captures hierarchical text structures but does not utilize the full-context integration, and transformer-based models such as *BERT*, *RoBERTa*, *ALBERT*, and *DeBERTa*, all fine-tuned with full-context integration. The transformer models incorporate claim-specific processing and external knowledge retrieval into their input data, enhancing their ability to detect misinformation by concatenating both the article's content and relevant external information.

D. Results and Analysis

The experimental results in Table I demonstrate that our *FCRV* model significantly outperforms all baseline models on both the GossipCop dataset and the out-of-domain datasets (*AI-Generated* and *PolitiFact*). On GossipCop, *FCRV* achieved an accuracy of 88.0% and an *F1* score of 89.4%, surpassing the best baseline, *DeBERTa*, which had 85.0% accuracy and an *F1* score of 85.5%; on the *AI-Generated* dataset, *FCRV* attained 86.4% accuracy and an *F1* score of 88.5%, outperforming baselines with accuracies ranging from 68.8% to 84.8%; and on *PolitiFact*, *FCRV* achieved 71.6% accuracy and an *F1* score of 82.8%, significantly higher than baselines whose accuracies ranged from 50.0% to 63.0%. Notably, directly applying *LLama 3.1* to detect fake news without full-context integration resulted in a lower accuracy of 71.6%, an *F1* score of 82.8%, and a high false positive rate (FPR) of 72.4%, indicating that even advanced language models struggle without full-context processing. These findings highlight the effectiveness of our *FCRV* model due to its full-context processing, which combines claim-specific extraction and external knowledge retrieval to enhance verification by identifying unsupported claims and fictitious entities.

V. CONCLUSION

Our preliminary results support the hypothesis that providing the model with comprehensive content for semantic analysis enhances its robustness and improves its performance in out-of-domain scenarios. By supplying the model with a full-context view, it appears better equipped to interpret nuanced relationships, which may contribute to its increased resilience across varied content types. Furthermore, early experiments suggest that the absence of corroborative evidence serves as a indicator for identifying *AI*-generated fake news. This finding indicates that models can potentially leverage gaps in available information as a distinctive feature to more accurately differentiate between authentic and fabricated content.

REFERENCES

[1] P. Ferragina and U. Scaiella, "Tagme: on-the-fly annotation of short text fragments (by wikipedia entities)," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, 2010, pp. 1625–1628.

[2] S. Gong, R. O. Sinnott, J. Qi, and C. Paris, "Fake news detection through graph-based neural networks: A survey," *arXiv preprint arXiv:2307.12639*, 2023.

[3] C. Malon, D. Shah, and J. Jiang, "FACTKB: Generalizable factuality evaluation using language models enhanced with factual knowledge," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 1023–1035.

[4] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big data*, vol. 8, no. 3, pp. 171–188, 2020.

[5] J. Su, T. Y. Zhuo, J. Mansurov, D. Wang, and P. Nakov, "Fake news detectors are biased against texts generated by large language models," *arXiv preprint arXiv:2309.08674*, 2023.

[6] Y. Wang, J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, O. M. Afzal, T. Mahmoud, G. Puccetti, T. Arnold *et al.*, "M4gt-bench: Evaluation benchmark for black-box machine-generated text detection," *arXiv preprint arXiv:2402.11175*, 2024.