An Explainable KG-RAG-Based Approach to Evidence-Based Fake News Detection Using LLMs

Jonathan John Thomas, Radu-Casian Mihailescu

Heriot-Watt University
jjt2002@hw.ac.uk, r.mihailescu@hw.ac.uk

Abstract

The advent of the Internet and social media has led to the rapid proliferation of fake news. Current state-of-the-art approaches for evidence-based fake news detection primarily utilize vector-based Retrieval Augmented Generation (RAG) systems. Recent studies have proposed RAG systems that outperform vector-based RAG systems by modeling the document store as a Knowledge Graph (KG). In this work, we investigated the performance of a KG-RAG-based approach for evidence-based fake news detection on the AVeriTeC dataset.

Introduction

Fake news refers to misinformation found in disseminated media, including news articles and social media posts. Fake news detection systems can be broadly classified into content-based and evidence-based approaches. Numerous content-based approaches that use machine learning and deep learning methods to detect fake news have been investigated (Kaliyar et al. 2020; Al-Yahya et al. 2021; Li et al. 2022). However, most of these approaches lack explainability and reliability. In contrast, evidence-based fake news detection systems offer some level of explainability, as they generally incorporate source citations and present retrieved evidence in a human-readable format.

To advance research in this field, the 2024 Automated Verification of Textual Claims (AVeriTeC) shared task (Schlichtkrull et al. 2024) invited proposals for systems that can perform evidence-based fake news detection on the provided AVeriTeC dataset (Schlichtkrull, Guo, and Vlachos 2024). In this study, we investigate a KG-RAG-based approach to evidence-based fake news detection on the AVeriTeC dataset.

Related Work

The AVeriTeC shared task is concerned with predicting the veracity of claims in the AVeriTeC dataset, classifying them with the labels *"Supported", "Refuted", "Not Enough Evidence"* or *"Conflicting Evidence/Cherry Picking"*. Additionally, a knowledge store for each claim, containing documents scraped from the Web, is provided to retrieve relevant evidence for verifying the claim.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

The top-performing AVeriTeC systems, InFact (Rothermel et al. 2024), HerO (Yoon et al. 2024), and AIC CTU system (Ullrich, Mlynář, and Drchal 2024) primarily utilize vector-based Retrieval-Augmented Generation (RAG) systems to retrieve evidence, conducting dense retrieval using a large embedding model.

More recently, Knowledge Graph-based RAG (KG-RAG) systems, such as HippoRAG 2, have been shown to outperform vector-based RAG systems, particularly in multi-hop question answering (Gutiérrez et al. 2025). They represent document stores as knowledge graphs (KGs) to better link information between disparate documents. Therefore, they show strong potential for effective evidence retrieval in fake news detection, particularly for complex claims requiring multi-hop questions to verify them. However, to the best of our knowledge, the evidence retrieval capabilities of KG-RAG systems on the AVeriTeC dataset have not been previously investigated.

Proposed Approach

The implemented KG-RAG-based framework for evidence-based fake news detection, depicted in Figure 1, performs the following steps:

1. Claim Type Classification: The claim type(s) (*'Position Statement', 'Event/Property Claim', 'Causal Claim', 'Numerical Claim', 'Quote Verification'*) are identified using a fine-tuned DeBERTaV3-large model (He, Gao, and Chen 2021; He et al. 2021).
2. Example Retrieval: BM25 (Robertson and Zaragoza 2009) is used to retrieve example claims from the training set that have the same claim type(s) as the input claim, along with their corresponding evidence questions. These few-shot examples will be provided to the Question Answering Large Language Model (Q&A LLM) in step 4 to guide question generation.
3. KG Construction (HippoRAG 2 Indexing): The system utilizes the HippoRAG 2 framework for indexing and retrieving evidence from each claim's knowledge store. During indexing, entities are first extracted using the gliner-medium-news-v2.1 model (Törnquist and Caulk 2024; Zaratiana et al. 2024). Next, relationship triples are extracted from the documents using a fine-tuned Llama-3.2-1B model (Meta 2024), trained for triple extraction

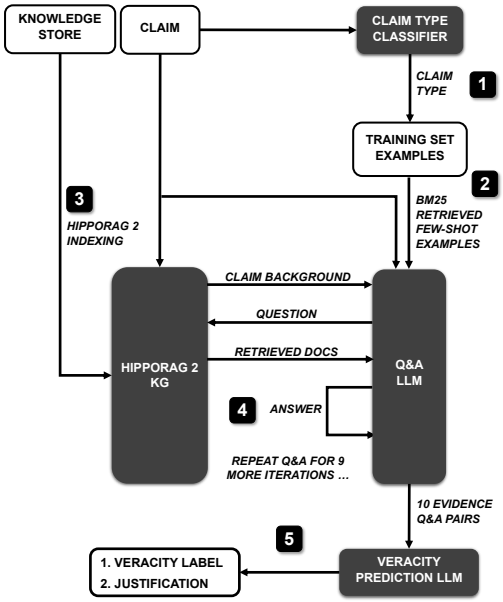

Figure 1: KG-RAG-Based Evidence-Based Fake News Detection Framework

through model distillation. The extracted entities and triples are added as nodes and edges in a KG, respectively, along with the documents themselves. Additionally, a retrieval encoder is used to create additional edges between semantically similar nodes/entities (Gutiérrez et al. 2025).

4. Multi-Step Evidence Retrieval: Inspired by Malon (2024), we employ a multi-step evidence retrieval strategy in our framework:
   (a) Prior to question generation, relevant background information is retrieved from the constructed KG using HippoRAG 2.
   (b) The claim, its metadata (claim date, speaker, original claim URL, claim reporting source, and the location ISO code relevant to the claim), the retrieved background information, and the example claims and questions from step 2 are provided to a Q&A LLM, which is prompted to generate a question for retrieving relevant evidence.
   (c) The HippoRAG 2 framework is used to retrieve k documents from the KG that are relevant to the question.
   (d) The Q&A LLM uses the retrieved documents to answer the question and generates a follow-up question to retrieve further evidence from the knowledge store.
   (e) Steps 4(c) and 4(d) are repeated until the Q&A LLM asks and answers 10 questions, which is the maximum number of questions that the AVeriTeC scoring function considers per claim (Schlichtkrull, Guo, and Vlachos 2024).
5. Veracity Prediction: Finally, a veracity prediction LLM predicts the veracity label (*"Supported", "Refuted", "Not Enough Evidence"* or *"Conflicting Evidence/Cherry Picking"*) using the 10 evidence Q&A pairs from step 4. Additionally, it is prompted to provide a justification, to leverage the benefits of chain-of-thought prompting (Wei et al. 2024).

System Name	Q only	Q + A	AVeriTeC Score
InFact	0.45	0.34	0.63
Baseline	0.24	0.20	0.11
KG-RAG Framework	0.43	0.31	0.32

Table 1: Evaluation results of the proposed KG-RAG framework, current state-of-the-art (InFact) and AVeriTeC baseline (Schlichtkrull et al. 2024). InFact and Baseline were evaluated on the full test set; KG-RAG framework on a subset of the development set.

Experimental Setup

The AVeriTeC score was used for evaluating the performance of the framework, and the Q only and Q+A scores were used to assess the quality of evidence retrieved by the system (Schlichtkrull, Guo, and Vlachos 2024). The framework was evaluated on 100 randomly selected claims from the AVeriTeC dataset's development set, sampled to ensure a balanced class distribution. This is because verifying a single claim took, on average, 11.86 hours with available hardware, due to the large size of the AVeriTeC knowledge stores (approximately 1000 documents for each claim).

Contriever (Izacard et al. 2022) was used as the retrieval encoder for HippoRAG 2 indexing and retrieval. Additionally, during each step of evidence retrieval, the top $k = 5$ documents returned by HippoRAG 2 were retrieved. Finally, Phi 4 (14.7B) (Abdin et al. 2024) and DeepSeek-R1-Distill-Qwen-32B (DeepSeek-AI 2025) were selected as the Q&A and veracity prediction LLMs, respectively.

Results and Analysis

The proposed KG-RAG framework attained an AVeriTeC score of 0.32, 0.21 greater than the baseline (see Table 1). Although the system does not achieve an AVeriTeC score higher than the current state-of-the-art system, InFact, it demonstrates strong performance using comparatively smaller, open-source LLMs - unlike InFact, which used closed-source models. Additionally, it attains Q only and Q+A scores comparable to InFact's, indicating the high quality of evidence retrieved by the KG-RAG framework.

Conclusion

In conclusion, this study proposed a KG-RAG-based framework for evidence-based fake news detection on the AVeriTeC dataset, achieving an AVeriTeC score of 0.32 on a subset of the development set. The primary limitation of the study was the computational challenge associated with indexing the large knowledge stores of the AVeriTeC dataset. Future work could evaluate the framework more thoroughly on the full AVeriTeC hidden test and experiment with more powerful LLMs and retrieval encoders to improve results.

References

Abdin, M.; Aneja, J.; Behl, H.; Bubeck, S.; Eldan, R.; Gunasekar, S.; Harrison, M.; Hewett, R. J.; Javaheripi, M.; Kauffmann, P.; Lee, J. R.; Lee, Y. T.; Li, Y.; Liu, W.; Mendes, C. C. T.; Nguyen, A.; Price, E.; de Rosa, G.; Saarikivi, O.; Salim, A.; Shah, S.; Wang, X.; Ward, R.; Wu, Y.; Yu, D.; Zhang, C.; and Zhang, Y. 2024. Phi-4 Technical Report. arXiv:2412.08905.

Al-Yahya, M.; Al-Khalifa, H.; Al-Baity, H.; AlSaeed, D.; Essam, A.; Uddin, M. I.; and Uddin, M. I. 2021. Arabic Fake News Detection: Comparative Study of Neural Networks and Transformer-Based Approaches. *Complexity (New York, N.Y.)*, 2021(1).

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Gutiérrez, B. J.; Shu, Y.; Qi, W.; Zhou, S.; and Su, Y. 2025. From RAG to Memory: Non-Parametric Continual Learning for Large Language Models. arXiv:2502.14802.

He, P.; Gao, J.; and Chen, W. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. arXiv:2111.09543.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations*.

Izacard, G.; Caron, M.; Hosseini, L.; Riedel, S.; Bojanowski, P.; Joulin, A.; and Grave, E. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. *Transactions on Machine Learning Research*.

Kaliyar, R. K.; Goswami, A.; Narang, P.; and Sinha, S. 2020. FNDNet – A Deep Convolutional Neural Network for Fake News Detection. *Cognitive systems research*, 61: 32–44.

Li, S.; Yao, T.; Li, S.; and Yan, L. 2022. Semantic-Enhanced Multimodal Fusion Network for Fake News Detection. *International journal of intelligent systems*, 37(12): 12235–12251.

Malon, C. 2024. Multi-hop Evidence Pursuit Meets the Web: Team Papelo at FEVER 2024. In Schlichtkrull, M.; Chen, Y.; Whitehouse, C.; Deng, Z.; Akhtar, M.; Aly, R.; Guo, Z.; Christodoulopoulos, C.; Cocarascu, O.; Mittal, A.; Thorne, J.; and Vlachos, A., eds., *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, 27–36. Miami, Florida, USA: Association for Computational Linguistics.

Meta. 2024. meta-llama/Llama-3.2-1B. https://huggingface.co/meta-llama/Llama-3.2-1B. Accessed: 2025-05-07.

Robertson, S.; and Zaragoza, H. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval*, 3(4): 333–389.

Rothermel, M.; Braun, T.; Rohrbach, M.; and Rohrbach, A. 2024. InFact: A Strong Baseline for Automated Fact-Checking. In Schlichtkrull, M.; Chen, Y.; Whitehouse, C.; Deng, Z.; Akhtar, M.; Aly, R.; Guo, Z.; Christodoulopoulos, C.; Cocarascu, O.; Mittal, A.; Thorne, J.; and Vlachos, A., eds., *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, 108–112. Miami, Florida, USA: Association for Computational Linguistics.

Schlichtkrull, M.; Chen, Y.; Whitehouse, C.; Deng, Z.; Akhtar, M.; Aly, R.; Guo, Z.; Christodoulopoulos, C.; Cocarascu, O.; Mittal, A.; Thorne, J.; and Vlachos, A. 2024. The Automated Verification of Textual Claims (AVeriTeC) Shared Task. In Schlichtkrull, M.; Chen, Y.; Whitehouse, C.; Deng, Z.; Akhtar, M.; Aly, R.; Guo, Z.; Christodoulopoulos, C.; Cocarascu, O.; Mittal, A.; Thorne, J.; and Vlachos, A., eds., *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, 1–26. Miami, Florida, USA: Association for Computational Linguistics.

Schlichtkrull, M.; Guo, Z.; and Vlachos, A. 2024. AVERITEC: A Dataset For Real-World Claim Verification with Evidence From The Web. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

Törnquist, E.; and Caulk, R. A. 2024. Model Card for gliner_medium_news-v2.1. https://huggingface.co/EmergentMethods/gliner_medium_news-v2.1. Accessed: 2025-04-10.

Ullrich, H.; Mlynář, T.; and Drchal, J. 2024. AIC CTU system at AVeriTeC: Re-Framing Automated Fact-Checking as a Simple RAG Task. In Schlichtkrull, M.; Chen, Y.; Whitehouse, C.; Deng, Z.; Akhtar, M.; Aly, R.; Guo, Z.; Christodoulopoulos, C.; Cocarascu, O.; Mittal, A.; Thorne, J.; and Vlachos, A., eds., *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, 137–150. Miami, Florida, USA: Association for Computational Linguistics.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2024. Chain-of-Thought Prompting Elicits Reasoning In Large Language Models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713871088.

Yoon, Y.; Jung, J.; Yoon, S.; and Park, K. 2024. HerO at AVeriTeC: The Herd of Open Large Language Models for Verifying Real-World Claims. In Schlichtkrull, M.; Chen, Y.; Whitehouse, C.; Deng, Z.; Akhtar, M.; Aly, R.; Guo, Z.; Christodoulopoulos, C.; Cocarascu, O.; Mittal, A.; Thorne, J.; and Vlachos, A., eds., *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, 130–136. Miami, Florida, USA: Association for Computational Linguistics.

Zaratiana, U.; Tomeh, N.; Holat, P.; and Charnois, T. 2024. GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 5364–5376. Mexico City, Mexico: Association for Computational Linguistics.