International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

AI-Based News Verification System Using Large Language Models and Retrieval-Augmented Generation

Aliraja Ansari *, Nakka Sharmila, Dasari Sahitya Lalitha Sri, Bhoga Bhadragiri and G. S. N. Murthy

Department of Computer Science and Engineering, Aditya College of Engineering and Technology, Surampalem, Kakinada, Andhra Pradesh, India.

International Journal of Science and Research Archive, 2026, 18(03), 513-522

Publication history: Received on 21 January 2026; revised on 05 March 2026; accepted on 07 March 2026

Article DOI: https://doi.org/10.30574/ijsra.2026.18.3.0471

Abstract

The pace with which digital news media and social media started proliferating has intensified the rate of misinformation, bringing significant problems with the reliability of information and the credibility of the people. Traditional fake news detection systems rely on the traditional way of machine learning systems, which are fed by predefined data sets, which restricts its flexibility to new events and real-time changes. Also, stand-alone large language models (LLMs) are likely to be susceptible to failing to base their responses on the existing evidence, which in turn leads to a high risk of hallucination and contextual bias. The aim of the paper is to suggest a real-time AI-based News Verification System that will be built by incorporating Retrieval-Augmented Generation (RAG) with a Large Language Model (Gemini 2.0 Flash) to achieve context-sensitive and explainable news content verification. It is founded on a modular architecture of a rest-based architecture written in FastAPI as a backend and Next.js as a frontend, MongoDB as a persistence layer and JWT as an authentication. The Tavily Search API retrieves real-time contextual evidence and then uses it together with the logic of LLM to ensure they become more credible and less groundless. The framework generates ordered output in terms of classification label (Real/Fake), credibility score (0-100%), summary of explanation and identification of suspicious phrases. Performance assessment identifies a mean of 2.8 seconds response latency time when the system is stable with simultaneous API requests. The suggested architecture offers an architecture which offers scalability, modularity, and production readiness to detect misinformation in real-time in dynamic digital environments.

Keywords: Fake News Detection; Retrieval Augmented Generation; Large Language Model; Real Time News Verification; Credibility Scoring; REST Based Architecture; Explainable Artificial Intelligence; Context Aware Artificial Systems

1. Introduction

The speed with which digital news outlets and social media are expanding has greatly contributed to the dissemination of fake news. Misleading or fake information can have a crucial impact on people in shaping their opinions, political stability, and affecting the health and economic systems. With the increasing volume of the online information, scalable and real-time verification systems have become necessary. Current fact-checking is based on human professionals and institutionalized structures that are not capable of keeping pace with the pace of contemporary information distribution. Classical machine learning-based fake news detection systems which are automated endeavors are designed to combat this problem, relying heavily on static, pre-annotated datasets and facing emerging or novel events. The recent developments in Large Language Models (LLMs) have superior contextual reasoning functions. Nevertheless, independent LLMs can produce answers without basing them on existing evidence, hallucinating or making outdated assumptions. Retrieval-Augmented Generation Retrieval-Augmented Generation (RAG) removes this drawback by incorporating external evidence retrieval in the inference mechanism, thus enhancing factual grounding. The present

* Corresponding author: Aliraja Ansari

paper proposes a News Verification System in the form of a real-time AI solution that combines RAG with a Large Language Model (Gemini 2.0 Flash) as part of a modular architecture based on a REST interface. Dynamic retrieval of contextual evidence, structured verifying outputs in the form of classification labels and credibility scores are generated, and a secure and scalable deployment is ensured by FastAPI, Next.js, MongoDB, and JWT-form based authentication. The suggested framework focuses on real-time flexible design, clarification and production capable system structure to combat misinformation in dynamic online space.

2. Literature Survey

Detection of fake news has been extensively investigated based on both conventional machine learning and deep learning methods. The initial systems were based on feature based representation like TF-IDF together with the classifier like Support Vector Machines and Naive Bayes [1], [2]. Although these methods proved to be reasonable in benchmark data, it relies on pre-labeled data and manual features. Consequently, they are limited with regard to their flexibility to new issues and changing misinformation trends. Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks became the subsequent deep learning models, which enhanced semantic comprehension and extraction of contextual features [3], [4]. These architectures offer superior language patterns in terms of sequential and hierarchical language patterns compared to traditional classifiers. Nevertheless, in spite of better representation learning, these models are still dataset-specific and can be of little use in terms of transparency in the real world. Large Language Models (LLM) have now proven to be capable of good contextual reasoning and understanding of language with the introduction of transformer-based architectures, most notably the model suggested by Vaswani et al. [11]. Such pre-trained systems like BERT [6] and GPT-based models [7], [9] have also promoted the performance of natural language processing in a broad range of tasks. Verification systems based on the LLM are in a position to process textual coherence and logical consistency in a better way than the traditional ones. However, untrained LLM mainly employs pre-training information and can give hallucinated or obsolete answers to new events. To overcome this shortcoming, Retrieval-Augmented Generation (RAG) was suggested to be a hybrid system that integrates the retrieval of external knowledge with the process of generative reasoning [8]. RAG obsoletes unfounded assertions and enhances grounding in facts by accessing pertinent evidence before actually generating responses. Similar datasets of fact-validation like FEVER [10] also underscore the relevance of evidence-based validation schemes. Although this has been provided, most of the current deployments have been made to put emphasis more on model performance in controlled dataset conditions than in scalability and production ready architecture. A gap still exists in creating a real-time, system-level framework that combines retrieval-augmented reasoning with a secure deployment that uses REST and persistent storage and explainable output generation. The proposed system fills this gap by having an integrated engineering based architecture of real time news verification.

3. System Architecture

The proposed News Verification System has a modular client-server architecture based on the principles of REST architecture, which is created in relation to real-time processing, scalability, and secure deployment. It consists of four major layers, which are Frontend Layer, Backend Layer, AI Processing Layer, and Database Layer.

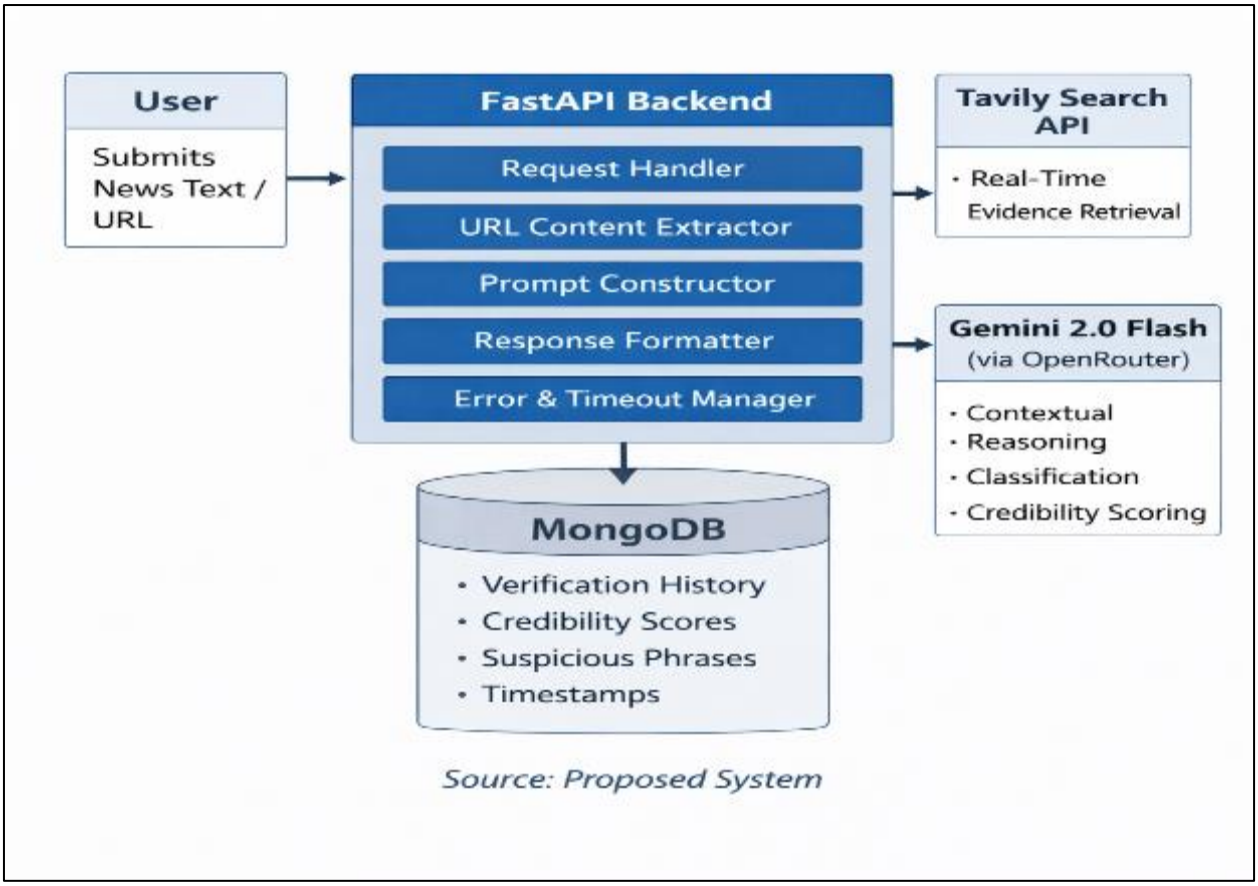

Figure 1 Proposed Real-Time News Verification System Architecture

3.1. Frontend Layer

The frontend is written on Next.js and offer an interactive user interface where news text or URLs should be submitted. It processes user input validation, category choice and organized visualization of outputs, such as classification labels, credibility ratings, summary of explanations and emphasis on suspicious phrases. The frontend interacts with the backend via secure calls using REST API.

3.2. Backend Layer

FastAPI is used to develop the backend which acts as the orchestration point. It handles the validation of requests, content of the URL, API routing and response formatting. When the user input is received, the backend is used to start the retrieval of evidence in real-time and build a structured prompt to the LLM. Error management, control of timeouts and processing multiple requests are also done in the backend in order to have a stable system performance.

3.3. AI Processing Layer

The AI layer is a combination of Retrieval -Augmented Generation (RAG) together with the Gemini 2.0 Flash Large Language Model using the OpenRouter gateway. The Tavily Search API allows retrieving real-time contextual evidence and it is passed in addition to the original news content. The LLM then processes the augmented input to produce structured outputs, including:

- Classification (Real/Fake)
- Credibility Score (0–100%)
- Explanation Summary
- Suspicious Phrase Identification

This argument based on retrieval minimizes all unsubstantiated assertions and maximizes the contextual fit.

3.4. Database and Persistence Layer

MongoDB is applied to persistent storage of verification history, user details, credibility ratings, timestamps and system logs. The database layer provides a guarantee of traceability and provides the future possibility to extend analytically. The API-level architecture is stateless whereas the storage-layer architecture is structured.

3.5. Security and Access Control

JWT authentication of secure endpoints is used to implement security. Environment variables are safely controlled, and policies of Cross-Origin Resource Sharing (CORS) are established in order to curb unauthorization. The REST architecture with no state improves scalability and secure sessions.

4. Methodology

The suggested system is based on a Retrieval-Augmented Generation (RAG)-workflow and it uses it to conduct verification of real-time news. The methodology incorporates input processing, contextual evidence retrieval, reasoning by LLM models, structured generation of output and persistent storage.

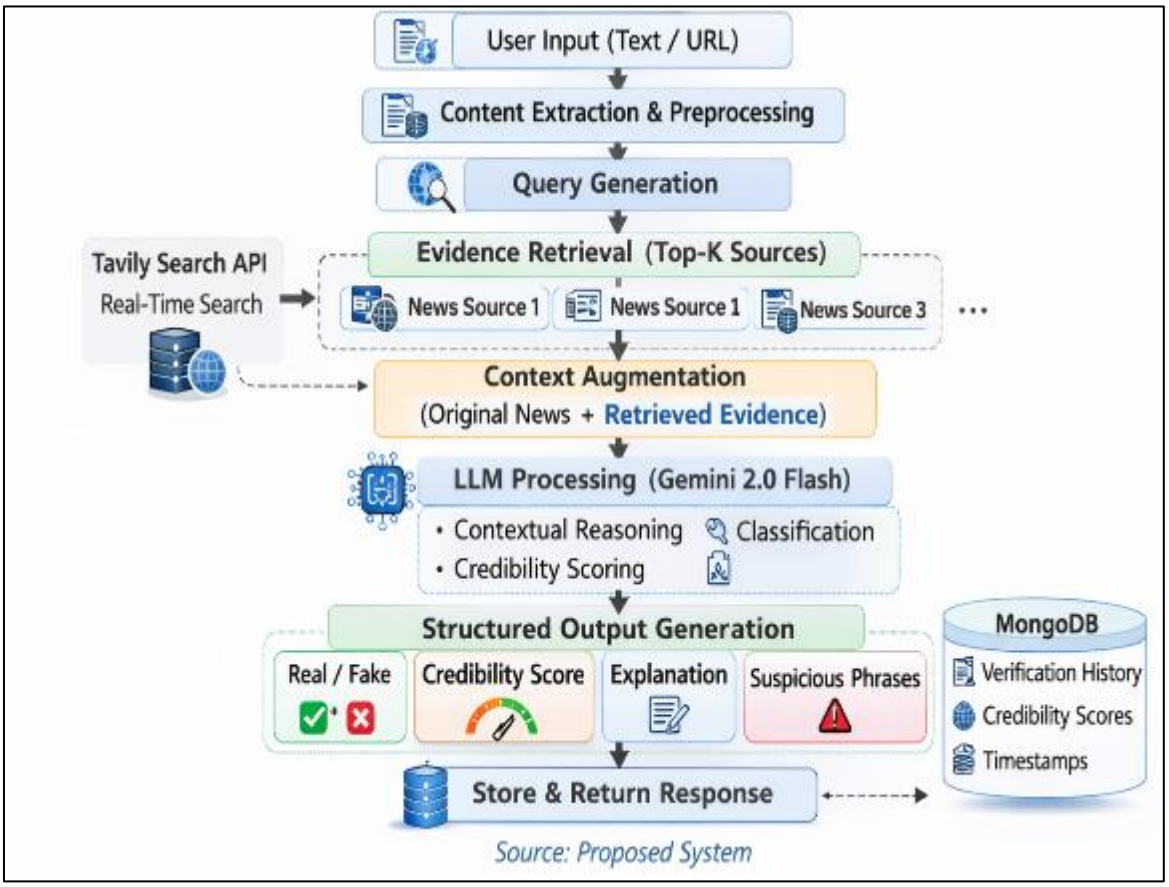

Figure 2 Retrieval-Augmented News Verification Workflow

4.1. Input Processing

The process starts with a user entering raw news text or a URL using the frontend interface and starts the verification process. In case of a URL is given, the backend just parses the article content through automated methods. The mined text is filtered and made standard to eliminate unwarranted metadata and then processed further.

4.2. Real-Time Evidence Retrieval

This is performed by retrieving evidence in real-time as it becomes accessible and is located. In order to improve factual grounding, the system retrives any of the relevant contextual evidence via the Tavily Search API. The backend produces

a query founded on the news material that is entered and gets the top-k sources that are relevant. These references give current background of an outsourced source, which is applied to build or refute the presented statement.

4.3. Prompt Construction and LLM Processing

The contextual evidence that has been retrieved is combined with the original news content to create a structured prompt. The augmented input is shipped to the Gemini 2.0 Flash model via OpenRouter gateway. The model does the contextual reasoning on the claim and the supporting evidence.

The LLM generates structured outputs including:

- Classification label (Real/Fake)
- Credibility score (0–100%)
- Explanation summary
- Identification of suspicious or inconsistent phrases

4.4. Output Structuring and Storage

The produced results are read in a structured form of a JSON and sent to the frontend to display. At the same time, data related to verification such as input text, credibility score, classification result, timestamp and metadata of the system are stored in MongoDB to ensure traceability and historical analysis.

4.5. Workflow Summary

The overall RAG-based verification workflow can be summarized as follows:

- Receive user input (text/URL).
- Extract and preprocess content.
- Retrieve real-time contextual evidence.
- Combine input and evidence into structured prompt.
- Perform LLM-based reasoning.
- Generate classification and credibility outputs.
- Store results and return response.

This structured workflow enables real-time, context-aware verification while maintaining scalability, modularity, and explainability.

5. Performance Evaluation and System Analysis

The proposed system was tested in terms of response latency, modular performance, concurrency processing, and stability of the entire system. This was an evaluation to determine real time feasibility and scalability under realistic deployment conditions.

5.1. Response Latency Analysis

The average end-to-end response time observed during testing was approximately **2.8 seconds** per verification request. The latency breakdown is as follows:

- Evidence Retrieval (Tavily API): 1.0–1.5 seconds
- LLM Inference (Gemini 2.0 Flash): < 1.0 second
- Backend Processing and Response Formatting: ~0.3–0.5 seconds

Figure 3 shows the latency distribution of key system parts. The findings show that real-time contextual retrieval is the main latency element, whereas LLM inference and the background processing are optimized. The system supports reasonable response time of interactive verification applications.

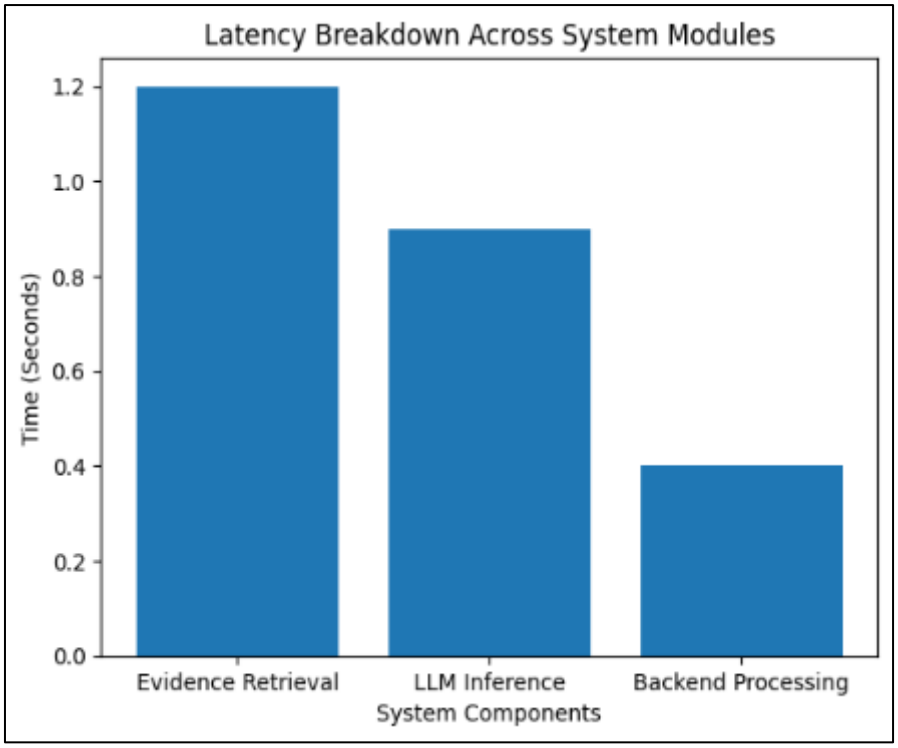

Figure 3 Latency Breakdown Across System Modules

5.2. Concurrent Request Handling

Scalability was tested by having moderate concurrent API calls to the backend. The stateless REST architecture and the asynchronous nature of request handling of FastAPI ensured that the system was not crashing or experiencing serious memory spikes, resulting in a stable performance. Latency of response was similar in simultaneous request conditions.

5.3. System Stability and Reliability

The module based architecture provides isolation among components minimizing the spread of failures. To avoid system breakdown, the backend has controlled exception handling and response management in the event of external API delay or timeout. Authentication of JWT and API key management of the environment are used to increase the security of operation. MongoDB is used as a persistent storage to implement traceability of verification history without impacting on API statelessness. Logging systems aid in debugging and monitoring in the deployment.

5.4. Engineering Observations

The combination of contextual grounding through retrieval and LLM reasoning enables better practical adaptability compared to static model-based systems. The architecture is also modular in nature to facilitate the addition of further retrieval sources or caching layers in the future.

The overall system exhibits desirable properties for scalability, latency, and production readiness in real-time misinformation verification environments.

6. Comparison with Existing Systems

Current fake news detection methods are based mainly on the conventional machine learning classifier, or independent deep learning and Large Language Model (LLM). Although these approaches are proven to be effective in the case of controlled data sets, most of them do not provide real-time flexibility, structured elucidation and production-deployable design. The conventional machine learning systems rely a lot on fixed datasets and defined features. The systems based on deep learning enhance contextual modeling but are data bound and computationally complex. Single LLM-based systems can improve the reasoning ability, but do not use retrieval to ground their model, which can lead to unsupported or hallucinated responses. Moreover, scalable deployment of REST and secure authentication systems as well as persistent management of verification history are not highlighted in many research implementations. The

suggested system is unique in its manner that Retrieval-Augmented Generation (RAG) is implemented in a modular, real-time and production-focused framework.

Table 1 Architecture-Level Comparison with Existing Approaches

Feature	**Traditional ML Systems**	**Standalone LLM Systems**	**Proposed RAG-Based System**
Real-Time Evidence Retrieval	✗	✗	✓
Adaptability to Emerging Events	Limited	Moderate	High
Contextual Grounding	Dataset-Dependent	Limited	Evidence-Grounded
Explainable Output Generation	Low	Moderate	High
Credibility Scoring	✗	Limited	✓
REST-Based Scalable Architecture	Limited	Limited	✓
Persistent Verification History	Rare	Rare	✓
Secure Authentication (JWT)	Rare	Rare	✓

The comparison highlights that while previous systems focus primarily on classification performance, the proposed framework emphasizes real-time contextual grounding, modular deployment, explainability, and secure engineering design. This system-level integration addresses practical deployment challenges often overlooked in model-centric research.

7. Results

7.1. Homepage Image

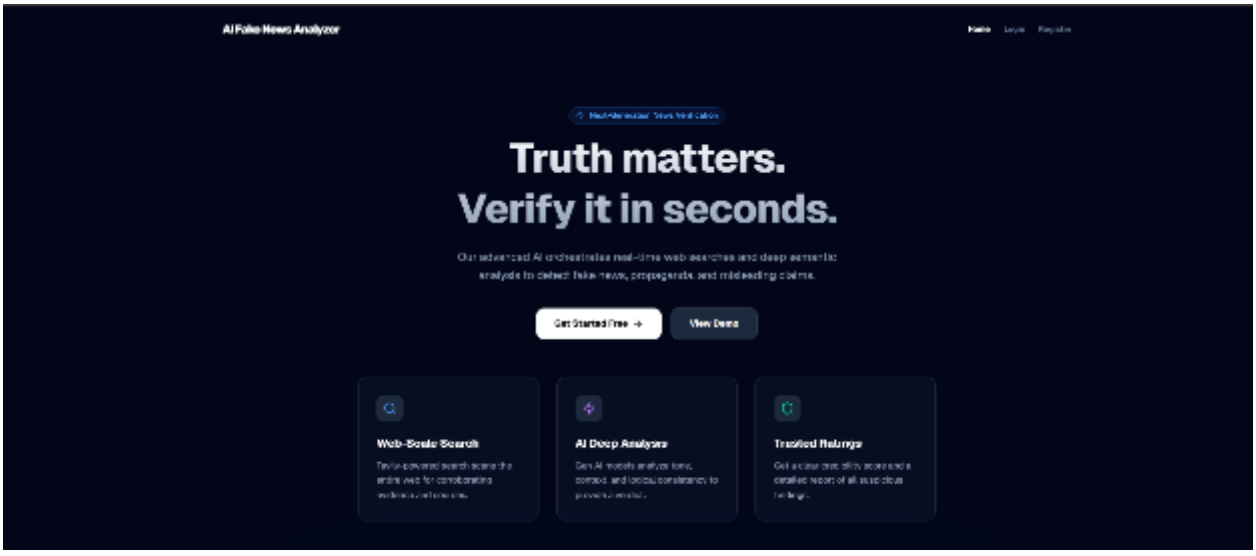

Figure 4 AI Based News Verification Homepage Interface

The homepage of this webpage is an AI Based News Verification System. Users are able to post news and the system scans the internet, analyzes the content using the AI, and informs whether the news is true or fake with a credibility rating. It operates through a web search, artificial intelligence analysis and reputation to check information in a fast manner.

7.2. Dashboard Image

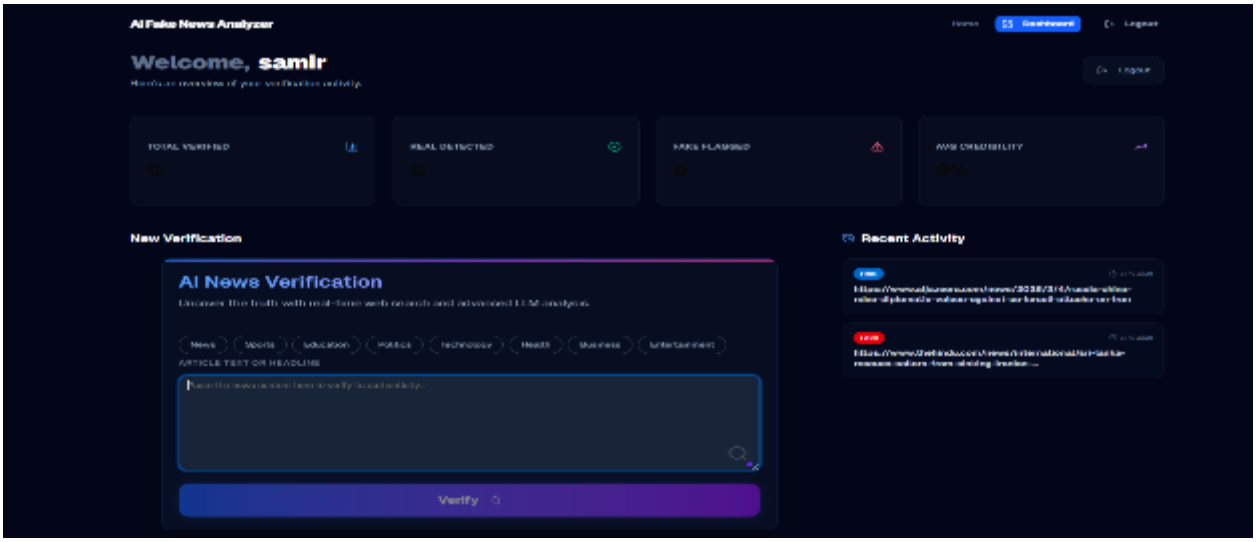

Figure 5 User Dashboard for News Verification Activity

In this dashboard the user is informed about his or her activity in terms of news verification real news detected, fake news flagged, and credibility scores. The user will have the option of pasting news article or headline and see whether it is a real or fake news with the help of AI. It also shows the recent check-ups so that the results are readily sought.

7.3. Verification Result Image

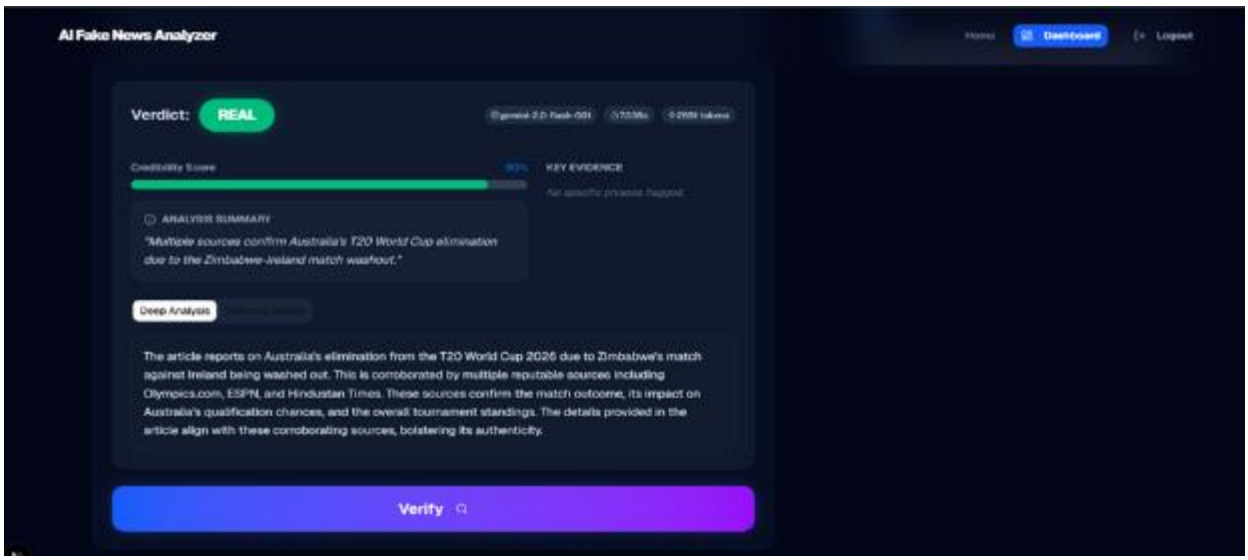

Figure 6 AI-Based News Verification Result with Credibility Score

In this screen, the AI verification of an article of news is presented. It shows whether the news is Real or Fake, the level of credibility and explanation based on evidence. The system is also able to give a justification of the decision by a thorough analysis by way of AI and web sources.

7.4. MongoDB Database Image

This screen is a database of MongoDB that holds the verification of news article results. The verification history collection stores information about the text/URL of the news, prediction label (Real/Fake), credibility score, and time. This enables the system to trace historical verifications and show them in the window of dashboard activity history.

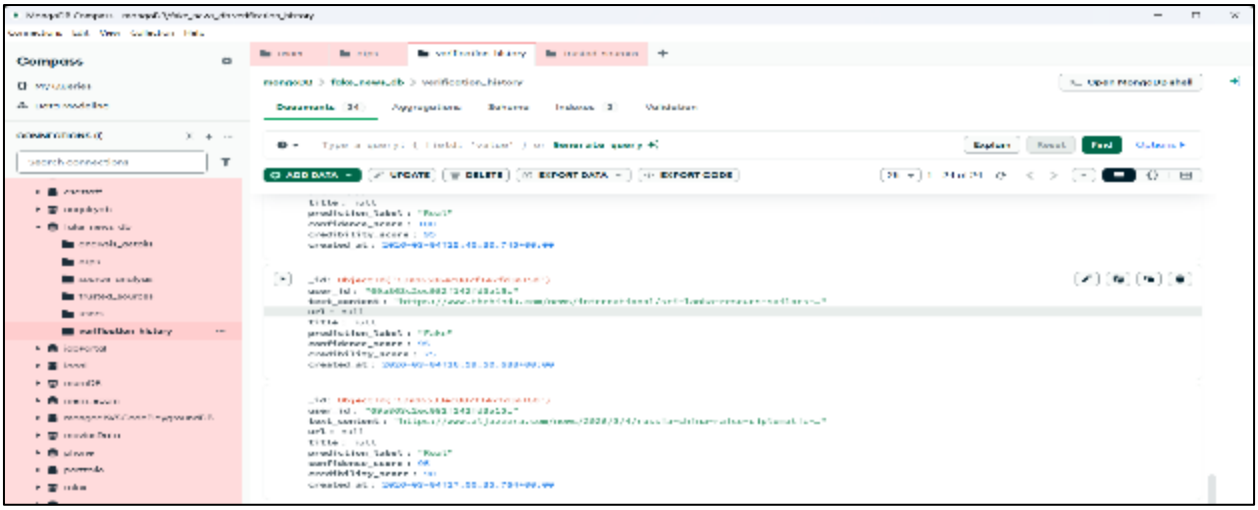

Figure 7 MongoDB Verification History Storage

7.5. Limitations

The proposed system has a number of limitations even though it is modular and scalable in nature. To begin with, the framework relies on the external APIs to retrieve evidence and make inferences using LLM. Response time and system availability may be impacted by network latency, API rate limits or service downtime. Third-party service reliability hence determines the performance of the verification process partially. Second, despite retrival-augmented generation minimizing the risks of hallucinations, it does not fully prevent the possibility of inaccuracy in reasoning by LLM. The score and classification outputs of credibility are still probabilistic and can be different based on the quality of evidence retrieved. Third, the system is currently working with English-language content as its main one. The support of multiple languages is not yet included, which restricts its application to non-English online space. Fourth, the quality of evidence in real-time is based on the indexing of search engines and credibility of content. When the sources used to be retrieved are biased or misleading the system can be reasoning accordingly. Lastly, the existing implementation focuses on system architecture and real-time deployment features as opposed to dataset-based benchmarking. Consequently, the comparisons of the accuracy of quantitative classification to the traditional models are beyond the limits of the present work.

7.6. Future Scope

The suggested system is a scalable base to real-time AI-based news verification, though, multiple improvements may enhance its power and strength. The next step can involve implementation of multilingual processing to aid in cross-lingual misinformation detection. The hybrid Retrieval-Augmented Generation could be enhanced by introducing a local vector database to decrease the reliance of external API and enhance the efficiency of retrieval. Also, the scale of high traffic settings can be improved with caching tools and implementation of distributed microservices. Structured bias detection modules and confidence calibration can also be used to enhance the credibility scoring mechanism to increase interpretability and fairness. Integration with verified fact-checking databases and knowledge graphs may also be a way to strengthen validation of the evidence. Lastly, extensive load analysis and testing in cloud-native would allow testing the framework in enterprise-scale traffic conditions, and the framework would become a complete production-grade misinformation monitoring system.

8. Conclusion

The present paper discussed a working verifying News AI-based system in real-time that combines Retrieval-Augmented Generation with a Large Language Model in a modular REST-based architecture. The suggested framework can overcome the critical drawbacks of the static, dataset-dependent methods of verification since it retrieves contextual evidence dynamically before the reasoning. The system is based on Fast API-based orchestration, Next.js based frontend interaction, MongoDB-based persistence and JWT-based authentication to guarantee scalability, modularity and secure deployment. Through a combination of evidences in real-time and LLM reasoning, the structure produces structured outputs such as classification labels, credibility scores, summaries of an explanation and suspicious phrases. The evaluation of performance shows the low-latency response properties and steady performance in the presence of concurrent requests, which means that real-time misinformation monitoring may become a viable solution. Although the present implementation focuses on the architecture of the system and readiness to deploy it, the modular

design can be extended to new features such as the use of multilingual text, hybrid retrieval system, and sophisticated calibration of the credibility. In general, the suggested framework is a scalable and production-focused approach to real-time, explainable news verification in dynamic online space.

Compliance with ethical standards

Acknowledgments

The authors would like to thank Aditya College of Engineering and Technology, Surampalem, for providing the academic environment and resources necessary to conduct this research.

Disclosure of conflict of interest

The authors declare that they have no conflict of interest.

References

[1] W. Y. Wang, "Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection," *Proc. ACL*, 2017.

[2] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake News Detection on Social Media: A Data Mining Perspective," *ACM SIGKDD Explorations*, vol. 19, no. 1, pp. 22–36, 2017.

[3] N. Ruchansky, S. Seo, and Y. Liu, "CSI: A Hybrid Deep Model for Fake News Detection," *Proc. CIKM*, 2017.

[4] Y. Liu and Y.-F. B. Wu, "Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks," *Proc. AAAI*, 2018.

[5] T. Baly et al., "Predicting Factuality of Reporting and Bias of News Media Sources," *Proc. EMNLP*, 2018.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proc. NAACL-HLT*, 2019.

[7] T. Brown et al., "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[8] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[9] OpenAI, "GPT-4 Technical Report," 2023.

[10] S. Thorne et al., "FEVER: A Large-Scale Dataset for Fact Extraction and Verification," *Proc. NAACL-HLT*, 2018.

[11] A. Vaswani et al., "Attention is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[12] M. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," *Proc. KDD*, 2016.

[13] J. Zhang and Y. Ghorbani, "An Overview of Online Fake News: Characterization, Detection, and Discussion," *Information Processing & Management*, vol. 57, no. 2, 2020.

[14] M. Allam and A. Dhunny, "On Big Data, Artificial Intelligence and Smart Cities," *Cities*, vol. 89, pp. 80–91, 2019.

[15] S. Raj and P. Bansal, "Scalable RESTful Architecture for Real-Time Web Applications," *International Journal of Computer Applications*, vol. 178, no. 40, 2019.