

# AI-Driven Explainable Fake News Detection Using Large Language Models, Evidence Retrieval and Source-Aware Reasoning

Tejal Bandu Gaikwad<sup>1</sup>, Mr. S. N. Shinde<sup>2</sup>, Prof. J. R. Mahajan<sup>3</sup>

Student, Department of Artificial Intelligence and Data Science<sup>1</sup>

Assistant Professor, Department of Artificial Intelligence and Data Science<sup>2</sup>

Head of Department (HOD), Department of Artificial Intelligence and Data Science<sup>3</sup>

Adsul Technical Campus, Chas, Ahilyanagar, India

**Abstract:** *The rapid growth of digital media has made false and misleading information easier to create and distribute. Earlier fake-news detection systems largely relied on machine-learning classifiers, text vectorisation, and ensemble voting. Recent work has introduced BERT, large language models (LLMs) and retrieval-augmented generation (RAG), but a practical detector still needs reliable evidence, source awareness, uncertainty handling and understandable explanations.*

*This paper proposes an AI-driven explainable framework that does not depend only on a trained classifier. The proposed system decomposes a news item into verifiable claims, retrieves supporting and contradicting evidence from multiple sources, evaluates source credibility and information freshness, performs semantic reasoning with an LLM, and combines the evidence into an evidence-weighted confidence score. Instead of forcing every input into a binary decision, the framework can return Real, Fake, or Uncertain together with the key evidence and a human-readable explanation. The paper presents the architecture, methodology, research gap, algorithm, evaluation plan and implementation design. The proposed approach is intended to improve transparency and robustness for both conventional misinformation and AI-assisted fabricated content.*

**Keywords:** Artificial Intelligence, Fake News Detection, Large Language Models, Retrieval-Augmented Generation, Natural Language Processing, Evidence Verification, Explainable AI, Source Credibility.

## I. INTRODUCTION

Online news can be produced and shared at very high speed. This creates a verification challenge because a misleading article may reach many users before conventional fact-checking is completed. The 2024 Procedia Computer Science study by Jouhar et al. investigates fake-news detection through data preprocessing, vectorisation, model training and machine-learning evaluation. The 2024 AntiFake System paper focuses on content-based automated verification. A 2025 Procedia Computer Science study further compares several classifiers, soft/hard voting ensembles, CNN and BERT. These studies demonstrate the value of automated classification, but their primary objective remains prediction from learned patterns. AI provides an opportunity to move from pattern-based classification toward evidence-based verification. LLM-based research has shown that language models can extract claims and reason over retrieved information. RAG-based systems can dynamically obtain external evidence, reducing dependence on a model's static training knowledge. However, LLMs can hallucinate and retrieval quality can vary, so evidence selection, source credibility and uncertainty are important design requirements. This research therefore proposes an AI-Driven Explainable Fake News Detection (ADEFND) framework. Its central idea is to treat a news article as a set of claims that must be checked against independent evidence rather than only asking a classifier whether the whole article looks



fake. The system produces a decision with supporting evidence and an explanation, making the output more useful for students, journalists, readers and content moderators.

## II. REVIEW OF EXISTING WORK

| Study                       | Main approach                                                         | Limitation/opportunity                                                     |
|-----------------------------|-----------------------------------------------------------------------|----------------------------------------------------------------------------|
| Jouhar et al. (2024)        | Machine-learning models, preprocessing, vectorisation, and evaluation | Primarily classification-oriented; limited evidence-based reasoning        |
| AntiFake System (2024)      | Content-based ML verification system                                  | Content is central; broader evidence fusion and explanation can be added   |
| Rahman et al. (2025)        | KNN, LR, DT, LGBM, XGBoost, RF, NB, voting, CNN and BERT              | Strong classification results; still mainly model-centric                  |
| LLM-based detection (2024)  | LLM/LLaMA-based disinformation detection                              | LLM knowledge can be stale; hallucination and verification remain concerns |
| RAG fact-checking (2024/25) | Claim extraction + retrieval + LLM reasoning                          | Retrieval quality and evidence reliability remain key challenges           |

Recent AI research is moving toward retrieval-enhanced and explainable fact verification. VeraCT Scan, for example, extracts core facts, searches for corroborating or conflicting reports and uses source credibility to support a transparent conclusion. Other 2025 work explores multi-source retrieval and redundancy filtering. These directions motivate the proposed combination of claim decomposition, source-aware evidence scoring, temporal freshness, uncertainty and explanation in one student-friendly architecture.

## III. RESEARCH GAP

Based on the reviewed studies, the following gaps are identified:

- **Binary decision limitation:** Real/fake output can be unsafe when available evidence is incomplete. An uncertainty state is useful.
- **Evidence reliability:** Retrieved information may be duplicated, outdated or low quality. Evidence should be weighted by source credibility and freshness.
- **Claim-level verification:** One article may contain several claims, some true and some false. Checking claims individually can provide a more informative result.
- **Explainability:** A prediction alone does not show why the system reached its conclusion. The proposed framework returns evidence and a concise rationale.
- **AI-generated misinformation:** Generative AI can create convincing fabricated text and entities. The detector should therefore check whether important claims have external support.
- **Static knowledge problem:** A model's internal knowledge can be incomplete or outdated. Retrieval from current sources is therefore included.

These gaps are consistent with recent literature highlighting retrieval quality, hallucination risk, evidence sufficiency and interpretability as important issues in LLM-based fake-news detection.

## IV. PROPOSED SYSTEM

The proposed ADEFND system combines Artificial Intelligence, Natural Language Processing, LLM-based semantic reasoning and Retrieval-Augmented Generation. Unlike a conventional classifier that maps an article directly to a class, the system first identifies claims, searches for evidence and then reasons over the evidence.



• Proposed AI-Driven Explainable Fake News Detection Architecture

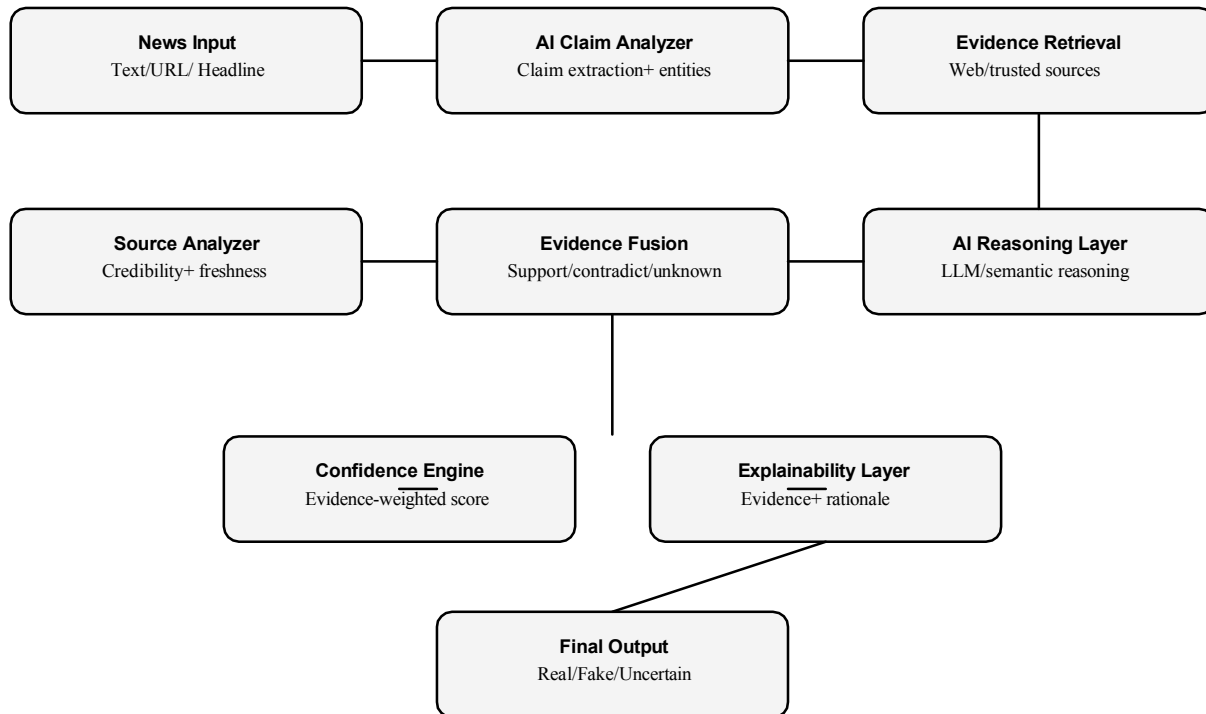


Figure 1. Proposed AI-driven architecture.

The architecture is a research design; the reported performance should be obtained through actual experiments and should not be claimed before implementation.

## V. MODULES OF THE PROPOSED SYSTEM

### 1.1 News Input Module

The user can enter a headline, article text or URL. The system normalises the input and removes irrelevant formatting while retaining important named entities, dates, locations and numerical facts.

### 1.2 AI Claim Extraction Module

An LLM/NLP model identifies the major factual claims in the article. For example, a sentence containing a date, person, organisation, and numerical statement can be converted into a structured claim such as: Entity – action – location – date – quantity. This enables claim-level verification.

### 1.3 Evidence Retrieval Module

For each claim, the system retrieves multiple independent evidence passages. Retrieval can use keyword search, semantic embeddings or a hybrid method. Evidence from multiple sources is retained so that the reasoning layer is not dependent on a single page.

### 1.4 Source Credibility and Freshness Module

Each evidence item receives metadata such as source type, publication date, update date, agreement with other sources and retrieval relevance. Recent authoritative evidence can receive a higher verification weight than an old or weakly supported source. The score is not intended to label a source as universally trustworthy; it is an input to the verification process.



### 1.5 AI Reasoning and Evidence Fusion

The LLM compares the claim with retrieved evidence and assigns one of three semantic relationships: supports, contradicts or insufficient evidence. The evidence-weighted results are then aggregated across claims.

### 1.6 Confidence and Explainability Module

Instead of outputting only a class, the system generates a confidence estimate and a short explanation containing the main verified claims, contradictory evidence and source references. If evidence is insufficient or contradictory, the system returns Uncertain rather than inventing a conclusion.

## VII. PROPOSED ALGORITHM

**Input:** News article N

**Output:** Label L {Real, Fake, Uncertain}, confidence C, explanation E

1. Receive news article N or a news URL.
2. Normalise and segment the text.
3. Use an AI language model to extract factual claims
4. For each claim, generate search/retrieval queries.
5. Retrieve candidate evidence from multiple sources.
6. Remove duplicate or highly redundant evidence.
7. Estimate relevance, source quality and temporal freshness for each evidence item.
8. Ask the reasoning model to classify each evidence–claim pair as Support, Contradict or Insufficient.
9. Aggregate weighted evidence across all claims.
10. If support is dominant with adequate evidence, assign Real; if contradiction is dominant, assign Fake; if evidence is weak or conflicting, assign Uncertain.
11. Generate an explanation containing the decisive claims and evidence.
12. Return label, confidence, explanation and evidence list.

A conceptual evidence score can be represented as: Score(C)

$$= [R \times Q \times F \times A]$$

where R is retrieval relevance, Q is source-quality weight, F is freshness weight, and A is agreement strength. The weights must be calibrated experimentally; they are not claimed as validated values in this paper.

## VIII. EXISTING SYSTEM VS PROPOSED SYSTEM

| Parameter            | Existing ML-oriented approach     | Proposed AI-oriented approach                    |
|----------------------|-----------------------------------|--------------------------------------------------|
| Core technique       | ML classifiers/ensembles/BERT     | LLM + RAG + semantic reasoning + evidence fusion |
| Input analysis       | Text features/embeddings          | Claim-level semantic decomposition               |
| Knowledge            | Mostly learned from training data | External retrieved evidence plus model reasoning |
| Verification         | Pattern/classification based      | Support vs. contradiction evidence checking      |
| Output               | Usually Real/Fake                 | Real/Fake/Uncertain + confidence                 |
| Explanation          | Often limited                     | Evidence-backed human-readable rationale         |
| Fresh information    | Not inherently dynamic            | Can retrieve current evidence                    |
| AI-generated content | May require additional training   | Can check claims against external evidence       |

The proposed design is not presented as a replacement for every classifier. Instead, it changes the verification objective from “Does this text look like fake news?” to “Which factual claims are supported or contradicted by available evidence, and how confident is that conclusion?”.



### **IX. DATASET AND EXPERIMENTAL PLAN**

For implementation, the study can use publicly available benchmark datasets such as LIAR, FakeNewsNet or other fact-checking datasets, subject to their licenses and task suitability. A separate evaluation set containing recent and AI-generated misinformation examples can be constructed where permitted. The dataset should be split into training/development/test sets without leakage between near-duplicate articles.

#### **Recommended experiments include:**

- Baseline 1: TF-IDF + Logistic Regression.
- Baseline 2: Random Forest / XGBoost or another conventional classifier.
- Baseline 3: BERT-based classifier.
- Proposed: LLM + RAG + evidence weighting + explanation.
- Ablation A: remove source-quality weighting.
- Ablation B: remove freshness weighting.
- Ablation C: remove claim decomposition.
- Ablation D: remove multi-source retrieval.

Metrics should include Accuracy, Precision, Recall, F1-score, confusion matrix and, for the three-state system, macro-F1. Because the proposed system outputs confidence, calibration metrics such as Expected Calibration Error can also be reported. For explainability, human evaluation can measure whether the cited evidence actually supports the generated rationale.

### **X. EXPECTED CONTRIBUTIONS**

The expected contributions of this research are:

1. A unified AI-based architecture for fake-news verification using LLM semantic reasoning and RAG.
2. Claim-level rather than only article-level verification.
3. Multi-source evidence fusion with source-quality and temporal-freshness factors.
4. An explicit Uncertain state to reduce forced decisions when evidence is insufficient.
5. Evidence-backed explanations to make the output understandable to users.
6. An experimental framework that compares conventional ML, BERT and the proposed AI pipeline.
7. A design that can later be extended to multimodal misinformation involving images, videos and AI-generated content.

### **XII. ADVANTAGES**

- More transparent than a prediction-only system because evidence can be displayed.
- Can use current retrieved information instead of relying exclusively on static model knowledge.
- Claim decomposition can identify which statement is problematic.
- Uncertainty handling can be used when sources disagree, or evidence is insufficient.
- The same architecture can be extended to different languages and domains after appropriate retrieval and model evaluation.

### **XIII. LIMITATIONS AND ETHICAL CONSIDERATIONS**

The proposed system should not be treated as an infallible truth oracle. Search engines and retrieved sources can contain errors, duplicated reporting and bias. LLMs can hallucinate, and a high confidence score does not guarantee factual correctness. Political, medical and emergency claims require especially careful source selection and human review. The system should present evidence and uncertainty rather than silently suppressing information. Evaluation should include false positives and false negatives because both can have serious consequences.



#### **XIV. FUTURE SCOPE**

Future work can extend the framework to multilingual Indian-language news, multimodal verification, image and video provenance, knowledge-graph evidence, cross-platform propagation analysis and real-time monitoring. A future version could combine retrieval with lightweight dense classifiers to reduce computational cost. Recent research suggests that multi-source retrieval, redundancy filtering and knowledge-graph-based RAG are promising directions for improving evidence-based verification.

#### **XV. CONCLUSION**

This paper proposes an AI-driven approach to fake news detection that moves beyond conventional machinelearning classification. The proposed ADEFND framework combines claim extraction, multi-source retrieval, source-aware evidence scoring, LLM reasoning, uncertainty handling and explainable output. The key research idea is to verify factual claims against evidence rather than relying only on linguistic patterns. This design is motivated by limitations identified in earlier ML-based systems and by recent progress in LLM and RAG-based fact verification. The framework provides a practical foundation for experimental implementation and comparison with existing approaches. Actual accuracy, precision, recall and F1-score must be obtained from controlled experiments; no performance value is fabricated in this paper.

#### **XVI. ACKNOWLEDGEMENT**

I would like to express my sincere gratitude to my project guide and faculty members of the Department of Artificial Intelligence And Data Science for their valuable guidance, continuous support, and encouragement throughout the preparation of this research paper. I am also thankful to my institute and Savitribai Phule Pune University for providing the necessary academic resources and learning environment. I would like to acknowledge the researchers and authors whose published work on fake news detection, Artificial Intelligence, Large Language Models, and evidence-based verification helped me understand the existing approaches and identify the research gap for the proposed system. Finally, I am grateful to my family and friends for their constant motivation, encouragement, and support during this work.

#### **REFERENCES**

- [1] J. Jouhar, A. Pratap, N. Tijo, and M. Mony, "Fake News Detection using Python and Machine Learning," *Procedia Computer Science*, vol. 233, pp. 763–771, 2024, doi: 10.1016/j.procs.2024.03.265.
- [2] S. Fedushko, Y. Syerov, and N. Kryvinska, "AntiFake System: Machine Learning-Based System for Verification of Fake News," *Procedia Computer Science*, vol. 238, pp. 663–670, 2024, doi: 10.1016/j.procs.2024.06.075.
- [3] A. Rahman et al., "Fake News Detection: Exploring the Efficiency of Soft and Hard Voting Ensemble," *Procedia Computer Science*, vol. 252, pp. 748–757, 2025, doi: 10.1016/j.procs.2025.01.035.
- [4] "Large Language Model Based Fake News Detection," *Procedia Computer Science*, vol. 231, pp. 740–745, 2024, doi: 10.1016/j.procs.2023.12.144.
- [5] Y. Bai and K. Fu, "A Large Language Model-based Fake News Detection Framework with RAG Fact-Checking," *IEEE International Conference on Big Data*, 2024, doi: 10.1109/BigData62323.2024.10826000.
- [6] C. Niu et al., "VeraCT Scan: Retrieval-Augmented Fake News Detection with Justifiable Reasoning," *ACL 2024 Demonstration Papers*, pp. 266–277, 2024, doi: 10.18653/v1/2024.acl-demos.25.
- [7] D. Li, F. Li, B. Song, L. Tang, W. Zhou, "IMRRF: Integrating Multi-Source Retrieval and Redundancy Filtering for LLM-based Fake News Detection," *NAACL-HLT 2025*, pp. 9127–9142, 2025, doi: 10.18653/v1/2025.naacl-long.461.
- [8] T. Zhao et al., "Generate First, Then Sample: Enhancing Fake News Detection with LLM-Augmented Reinforced Sampling," *ACL 2025*, pp. 24276–24290, 2025, doi: 10.18653/v1/2025.acl-long.1182.



[9] A. M. Qazi, J. P. McCrae, and J. A. Nasir, "When retrieval outperforms generation: Dense evidence retrieval for scalable fake news detection," LDK 2025, 2025

[10] J. J. Thomas and R.-C. Mihailescu, "An Explainable KG-RAG-Based Approach to Evidence-Based Fake News Detection Using LLMs,"

Proceedings of the AAAI Symposium Series, vol. 6, pp. 152–154, 2025, doi: 10.1609/aaaiss.v6i1.36047.

