Sushant Khanderao Sonawane, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

AI-Powered Real-Time Crisis Misinformation Debunker

Sushant Khanderao Sonawane, Ritik Prakash Patil, Shriraj Suresh Mahajan, Deepali S. Suryawanshi
Department of Artificial Intelligence and Data Science MET Institute of Engineering, Nashik, India

Abstract- The rapid spread of misleading information across digital platforms has become a significant concern, especially during critical situations such as public health emergencies, elections, and natural disasters. Traditional fact-checking ap- proaches, which depend largely on manual verification, are often too slow to respond effectively to the speed at which information circulates online. This limitation creates a need for automated systems capable of analyzing and verifying information in real time. This research presents an AI-based framework designed to identify and verify potentially false information efficiently. The proposed system combines multiple technologies, including Large Language Models (LLMs), Retrieval-Augmented Gener- ation (RAG), and Optical Character Recognition (OCR), to process both textual and visual content. By retrieving relevant information from trusted sources and analyzing it using intel- ligent models, the system generates clear and evidence-based conclusions along with confidence scores. In addition, the framework incorporates a self-learning knowl- edge structure that improves performance over time by storing validated information and learning from previous analyses. The system is also designed to provide understandable explanations for its decisions, increasing transparency and user trust. Overall, the proposed approach offers a scalable and practical solution for real-time misinformation detection across various digital platforms.

Keywords— Misinformation Detection, Large Language Models, Retrieval-Augmented Generation, Optical Character Recognition, Crisis Management, Explainable AI, Knowledge Graphs

I. INTRODUCTION

The advancement of digital communication technologies has significantly changed how information is created, shared, and consumed. Social media platforms, online news portals, and instant messaging services enable information to spread across the world within seconds. While this has improved connectivity and access to knowledge, it has also led to the rapid circulation of false or misleading information.

During critical situations such as pandemics, elections, and natural disasters, the impact of misinformation becomes more

severe. Incorrect or manipulated information can create panic among the public, influence decision-making, and reduce trust in reliable sources. Managing such situations requires accurate and

timely verification of information, which is difficult to achieve using traditional methods.

Conventional fact-checking processes rely on human experts to verify claims, making them reliable but slow. These methods are not suitable for handling the large volume and high speed of data generated in today's digital environment. As a result, there is a growing need for automated systems that can analyze information quickly and provide reliable results.

Recent developments in Artificial Intelligence, particularly in Natural Language Processing and deep learning, have made it possible to build systems that can understand and analyze textual data more effectively. However, many existing

solutions focus only on text-based inputs and often fail to handle information present in images or multimedia formats. Additionally, some systems lack the ability to provide clear explanations for their decisions, which reduces user trust.

To overcome these challenges, this research proposes an AI- powered system designed to detect and verify misinformation in real time. The system integrates multiple technologies, including Large Language Models for understanding context, Retrieval-Augmented Generation for accessing reliable infor- mation, and Optical Character Recognition for processing text within images. By combining these components, the system aims to provide accurate, transparent, and scalable solutions for misinformation detection.

The main objective of this work is to develop a system that not only identifies false information but also explains the reasoning behind its decisions. Such a system can be useful in applications like social media monitoring, public awareness systems, and crisis management platforms.

II. PROPOSED CORE MODULES

The AI-Powered Real-Time Crisis Misinformation De- bunker is designed as a modular and flexible system that brings together multiple AI techniques into a unified workflow. It is capable of processing different forms of input such as text, URLs, and images, and produces results that are not only accurate but also supported with explanations and evidence.

Input Acquisition and Preprocessing Module
This stage serves as the starting point of the system, where user inputs are collected in formats like text, images, or web links. Once the input is received, several preprocessing steps are applied to refine the data.

These steps include filtering noise, removing duplicate en- tries, and extracting important metadata. For images, Optical Character Recognition (OCR) using tools such as Tesseract and OpenCV is applied to convert visual text into machine- readable form. The system also incorporates language detec- tion and translation features to support multilingual content.

Classification and Semantic Analysis Module
After preprocessing, the refined text is analyzed using transformer-based models such as BERT, RoBERTa, or other fine-tuned language models. These models help in capturing the meaning and context of the content more effectively.

Based on the analysis, the system categorizes the input as authentic, suspicious, or fabricated. It also assigns a confidence value to indicate how reliable the prediction is. Additional operations include identifying key terms, analyzing sentiment, and generating contextual embeddings for further processing.

RAG-Based Evidence Verification Module
In this stage, the system evaluates the correctness of the in- formation using a Retrieval-Augmented Generation approach. Relevant data is first fetched from reliable sources such as fact- checking services, news platforms, and internal databases.

The collected evidence is then compared with the input using a language model, which produces a final decision along with supporting references. If external sources are unavailable, the system continues functioning by relying on its internal knowledge base.

Knowledge Graph and Self-Learning Module
This part of the system focuses on continuous improvement by maintaining structured information about verified claims, source credibility, and entity relationships.

As the system processes more inputs, the stored knowledge is updated, allowing it to improve prediction accuracy over time. This mechanism also reduces reliance on external APIs and enhances system efficiency.

Explainable Interface and Feedback Integration
The results are presented to the user in a clear and in- terpretable manner. Along with the classification output, the system provides confidence scores,

supporting evidence, and highlights relevant parts of the input.
The interface is designed for usability and can be extended to web and mobile platforms using frameworks such as Next.js and Node.js. Feedback provided by users is also incorporated to improve system performance over time.

Background

Growth of Misinformation in Digital Platforms

The increasing use of internet-based platforms has sig- nificantly changed how information is distributed and con- sumed. Social media, messaging applications, and online news platforms allow information to spread rapidly across large audiences. While this improves accessibility, it also contributes to the faster spread of misleading or false information.

During critical situations such as health crises, elections, or natural disasters, misinformation can create confusion, influence public behavior, and reduce trust in reliable sources. Traditional verification approaches depend on manual review, which is accurate but not suitable for handling large-scale, real-time data. This has created a need for automated systems capable of efficient and fast verification. [7]

Early Approaches Using Machine Learning

Initial solutions for misinformation detection relied on basic Natural Language Processing combined with machine learning algorithms. Techniques such as Naive Bayes, Support Vector Machines, and Logistic Regression were used to classify con- tent based on textual features like word patterns and frequency. Although these methods were effective for simple tasks, they were limited in understanding deeper context and se- mantic meaning. They also faced challenges with complex language features such as sarcasm, ambiguity, and multilingual content, making them less effective in real-world scenarios.

Deep Learning and Advanced Language Models

The adoption of deep learning improved the ability of sys- tems to analyze text more effectively. Models like Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) enabled better representation of textual patterns.
Further advancements with transformer-based models such as BERT and RoBERTa enhanced contextual understand- ing, leading to improved detection accuracy. However, these models require significant computational resources and may occasionally generate incorrect outputs. [8]

Multimodal and Retrieval-Based Techniques

Recent developments focus on combining different data formats to improve detection performance. Many misleading posts include visual elements such as images and screenshots along with text. OCR techniques help extract text from such visuals, enabling further analysis.

Another approach involves Retrieval-Augmented Genera- tion, where relevant information is first gathered from trusted sources and then processed by an AI model to generate responses. This method improves reliability by ensuring that outputs are supported by verified data rather than only gener- ated content. [3]

III. EXISTING APPROACHES / LITERATURE SURVEY

Traditional NLP and Machine Learning Methods

Initial efforts in misinformation detection mainly used Natu- ral Language Processing combined with conventional machine learning models. Techniques such as Naive Bayes, Support Vector Machines, and Logistic Regression were applied to classify textual data from news articles and social media using features like word frequency, sentiment, and syntactic structure.

These approaches relied on manually crafted features such as n-grams, TF-IDF scores, and part-of-speech tagging. While they worked reasonably well on smaller or controlled datasets, they lacked the ability to capture deeper contextual meaning. As a result, they struggled with complex linguistic patterns such as sarcasm, ambiguity, and multilingual content. Their limited adaptability to evolving online content and lack of real- time

capability reduced their effectiveness in crisis scenarios.

Deep Learning and Transformer-Based Approaches

The use of deep learning significantly enhanced misinforma- tion detection by improving contextual understanding. Models such as Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks enabled better representation of sequential and contextual information.

Further improvements were achieved with transformer- based architectures like BERT, RoBERTa, and XLNet, which capture long-range dependencies within text. These models de- liver high accuracy and can identify subtle language patterns. However, they require considerable computational resources and may sometimes generate outputs that appear correct but are factually inaccurate.

Multimodal and Retrieval-Based Systems

To address the limitations of text-only analysis, multimodal approaches were introduced that combine textual, visual, and metadata information. OCR techniques make it possible to extract text from images such as memes and screenshots, enabling analysis of visually shared misinformation.

Retrieval-based systems enhance verification by using ex- ternal data sources such as fact-checking services, news repositories, and online databases. These systems compare input content with verified information to improve reliability. However, their dependence on external services can lead to issues such as limited availability, incomplete coverage, and delays.

LLM and RAG-Based Hybrid Systems

Recent advancements focus on combining Large Language Models with Retrieval-Augmented Generation techniques. This hybrid approach leverages the reasoning capabilities of language models along with evidence retrieved from trusted sources.

In this setup, relevant data is first gathered from reliable databases and knowledge repositories. The language model then processes this information to generate responses that are
both accurate and explainable. Compared to earlier methods, this approach offers improved adaptability, better contextual understanding, and greater transparency. The addition of OCR further extends its ability to handle both textual and visual misinformation.

Overall, the LLM and RAG-based approach provides an effective solution for real-time misinformation detection, of- fering scalability, evidence-based reasoning, and improved reliability in critical scenarios.

Table 1: Comparative Analysis Of Existing Methodologies

Approach	Strengths	Limitations	Crisis Suitability height Traditional ML
Simple implementation	Limited contextual understanding	Poor Deep Learning	Improved context handling
Risk of hallucination	Moderate Multimodal	Combines text and image analysis	Dependency on external APIs
Good LLM + RAG (Proposed)	Evidence-driven reasoning	Complex system design	Excellent height

The LLM and RAG-based approach demonstrates strong suitability for crisis scenarios by providing reliable, evidence-supported, and explainable outputs.

IV. SYSTEM ARCHITECTURE

The proposed AI-Powered Real-Time Crisis Misinformation De- bunker is designed using a modular and scalable architecture that combines multiple AI components within a unified system. The framework is capable of processing different types of inputs such as text, URLs, and images, and produces results that are supported by evidence and clear explanations.

Input Acquisition and Preprocessing Module

This module serves as the entry point of the system and accepts user inputs in the form of text, images, or links. Once the input is received, preprocessing operations are applied to enhance data qual- ity. These operations include removing noise, eliminating duplicate content, and extracting useful metadata.

For image inputs, Optical Character Recognition (OCR) techniques using tools like Tesseract and OpenCV are applied to extract text from images such as memes and screenshots. The system also supports multiple languages through language detection and translation, enabling it to handle region-specific content effectively.

Classification and Semantic Analysis Module
After preprocessing, the data is analyzed using transformer-based models such as BERT, RoBERTa, or fine-tuned language models. These models help in understanding the context and meaning of the input content.

Based on this analysis, the system classifies the information into categories such as authentic, suspicious, or fabricated. It also provides a confidence score to indicate prediction reliability. Additional steps include keyword extraction, sentiment analysis, and generation of contextual representations for further processing.

RAG-Based Evidence Verification Module
This module performs the task of verifying the correctness of the input information. It uses a Retrieval-Augmented Generation approach, where relevant data is first collected from trusted sources such as fact-checking APIs, news platforms, and internal databases.

The retrieved information is then compared with the input using a language model, which produces a final decision along with supporting evidence. If external sources are not accessible, the system relies on internally stored data to ensure continuous operation.

Knowledge Graph and Self-Learning Module
This component helps the system improve over time by storing ver- ified information, source credibility data, and relationships between different entities in a structured format.

As more data is processed, the knowledge base is updated con- tinuously, allowing the system to make more accurate predictions. This also reduces dependency on external APIs and enhances system efficiency.

Explainable Interface and Feedback Integration
The output of the system is presented in a clear and understandable format. Along with the classification result, the system provides confidence scores, supporting evidence, and highlights important parts of the input used in decision-making.

The interface can be extended to web and mobile platforms using technologies such as Next.js and Node.js. User feedback is also collected and used to improve system performance over time.

System Architecture And Data Flow
The system follows a layered architecture that ensures scalability, modularity, and ease of understanding:

Table 2: System Architecture Layers

Layer	Components	Function	Technology heightInput
Text, URL, Image	Data ingestion	REST API, OCR Preprocessing	Cleaning, Tokenization
Normaliz ation	NLTK, spaCy Classification	BERT / RoBERTa	Initial labeling
Transfor mers RAG	Evidence retrieval	Fact verification	Vector DB, LLM Knowledge
Graph storage	Learning	Neo4j, FAISS Output	UI, Explanations
User interactio n	Next.js, XAI height		

Mathematical Framework
The system uses a mathematical framework that includes similarity measurement, confidence scoring, and classification thresholds to ensure reliable misinformation detection.

Text Similarity Assessment: The Jaccard similarity coeffi- cient is used to measure the similarity between the input claim and retrieved evidence:
$|A \cap B|$

Ensemble Classification: The final classification is obtained by combining AI-based and algorithm-based scores:

F = 0.6 × AIScore + 0.4 × AlgoScore (3)

Classification criteria:

- F < 30 → Fabricated
- 30 ≤ F < 60 → Suspicious
- F ≥ 60 → Authentic

Source Credibility Scoring: Source reliability is evaluated using predefined scores:

- Fact-checking organizations: 0.95
- Trusted news sources: 0.85
- Unverified sources: 0.50
- Unreliable sources: 0.15

V. ANALYSIS AND DISCUSSION

This section provides a detailed comparison of different misinfor- mation detection techniques and highlights the strengths of the pro- posed AI-based framework. The analysis considers key factors such as accuracy, adaptability, interpretability, and real-time performance. It also demonstrates how systems based on Large Language Models and Retrieval-Augmented Generation offer better performance com- pared to traditional and deep learning methods. The integration of multimodal processing and self-learning knowledge graphs further improves system reliability while reducing reliance on external APIs.

Comparative Performance Evaluation

A comparison of various misinformation detection approaches shows clear differences in their performance. Traditional machine learning methods such as Support Vector Machines and Naive Bayes rely heavily on manually designed features, which limits their ability to generalize across diverse datasets. Deep learning models like CNN and LSTM improve contextual understanding but lack strong mechanisms for factual verification.

Transformer-based models provide advanced language understand- ing but require high computational resources and may sometimes generate inaccurate outputs. The proposed hybrid approach com- bining RAG and LLM addresses these challenges by incorporating real-time evidence retrieval, improving both accuracy and contextual understanding. [6]

$$J(A, B) = \frac{|A \cup B|}{|A \cup B|} \quad (1)$$

Multimodal System Assessment

where A and B represent sets of tokens from the claim and evidence respectively.

Threshold values include:

- Deduplication threshold: 0.80
- Evidence relevance threshold: 0.30

2) Weighted Confidence Aggregation: The overall confidence score is calculated by combining multiple factors using weighted aggregation:

$$S = 0.40 \times AI + 0.25 \times Src + 0.15 \times Time + 0.10 \times Ref + 0.10 \times Spec$$

Multimodal systems enhance misinformation detection by com- bining analysis of text and visual content. The use of OCR enables extraction of text from images such as memes and screenshots, allowing the system to process a wider range of data. However, these systems may face challenges when dealing with low-quality images, multiple languages, or distorted text. Despite these limitations, multimodal approaches are essential for effective misinformation detection due to the increasing use of visual content on digital platforms. [5]

Real-Time Operational Considerations

Weight distribution:

- AI model confidence (40
- Source credibility (25
- Temporal consistency (15
- Cross-reference frequency (10
- Claim specificity (10

The final score is normalized between 0 and 100.

(2) Detecting misinformation in real time involves challenges related to speed, scalability, and reliability. Systems that depend only on external APIs or static datasets may face issues such as delays or unavailability of data sources.

The proposed framework addresses these issues by using multiple data sources along with fallback mechanisms. This ensures continu- ous operation even when external services are not accessible,

making the system more reliable for real-time applications.

Framework Advantages

The proposed system offers several advantages by combining intelligent decision-making with transparent outputs and scalable architecture. The integration of LLMs with RAG enables accurate and explainable results by using both reasoning and verified evidence.

The use of a self-learning knowledge graph reduces dependency on external sources and improves system performance over time. Addi- tionally, the modular design—covering OCR processing, multilingual support, evidence retrieval, and explainable output—makes the sys- tem suitable for large-scale deployment in real-world applications.

VI. CONCLUSION

The rapid spread of misinformation during critical situations such as pandemics, elections, and natural disasters has become a major global concern, requiring effective and reliable solutions. This research introduced an AI-based framework for real-time misinfor- mation detection by integrating Large Language Models, Retrieval- Augmented Generation, and Optical Character Recognition to handle both textual and visual data.

Compared to traditional methods, the proposed system combines contextual understanding with evidence-based verification, resulting in improved accuracy and transparency. The inclusion of a self-learning knowledge graph and multi-source data retrieval reduces de- pendency on external APIs while maintaining real-time performance. Furthermore, features such as explainable outputs, multilingual support, and evidence visualization enhance user trust and system usability. The framework has strong potential for deployment in applications such as social media monitoring, web platforms, and intelligent information systems.

Future work may include extending the system to detect deep- fake videos using multimodal learning techniques, implementing blockchain-based mechanisms for secure verification, developing adaptive learning models for continuous improvement, and applying federated learning to ensure data privacy across distributed systems. These enhancements will further strengthen the system as a scalable and reliable solution for combating misinformation globally.

REFERENCES

1. K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake News Detection on Social Media: A Data Mining Perspective," ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22–36, 2017.
2. W. Y. Wang, "Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), 2017.
3. J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: A Large-scale Dataset for Fact Extraction and Verification," in Proceed- ings of NAACL-HLT, 2018, pp. 809–819.
4. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, et al., "Retrieval- Augmented Generation for Knowledge-Intensive NLP Tasks," in Ad- vances in Neural Information Processing Systems (NeurIPS), 2020.
5. S. Zhang, K. Zhao, and M. Chen, "Explainable Fake News Detection: Towards Interpretable and Trustworthy AI," Information Processing & Management, vol. 59, no. 2, pp. 102123, 2022.
6. A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, "TweetCred: Real-time Credibility Assessment of Content on Twitter," in Social Informatics (SocInfo), 2014, pp. 228–243.
7. I. Segura-Bedmar, D. Mart´ınez, and M. Revert, "Multimodal Fake News Detection: A Survey," Information, vol. 13, no. 10, pp. 463–478, 2022.
8. X. Zhou and R. Zafarani, "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities," ACM Computing Surveys (CSUR), vol. 53, no. 5, pp. 1–40, 2020.
9. Y. Liu, J. Ott, N. Goyal, et al., "RoBERTa: A Robustly Optimized BERT Pretraining

Approach," arXiv preprint arXiv:1907.11692, 2019.
10. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of NAACL-HLT, 2019, pp. 4171–4186.
11. S. A. Qureshi, F. Iqbal, and T. H. Meraj, "Hybrid Machine Learning Model for Fake News Detection using Text and Image Fusion," IEEE Access, vol. 10, pp. 68921–68933, 2022.
12. R. B. Saha and D. Dey, "Combating Misinformation with Explainable AI: A Multimodal Framework," Journal of Intelligent Information Systems, vol. 61, no. 3, pp. 789–808, 2023.
13. D. Jain, M. Singh, and P. Bhardwaj, "Blockchain-based Framework for Transparent Fact Verification in Social Media," IEEE Internet of Things Journal, vol. 10, no. 4, pp. 2892–2903, 2023.
14. K. N. Singh and V. Sharma, "A Multilingual Fake News Detection System using OCR and NLP," Procedia Computer Science, vol. 218, pp. 182–190, 2023.
15. M. Srivastava, A. Patra, and P. Prakash, "Towards Robust Crisis Misinformation Detection using LLMs and RAG," arXiv preprint arXiv:2402.08319, 2024.