International Journal of Engineering and Techniques - Volume 12 Issue 2, March-April - 2026

RESEARCH ARTICLE OPEN ACCESS

# FAKE NEWS AND MISINFORMATION DETECTION USING NLP WITH LLM

Rajarajan.M[1], Azhara.M[2], Kaviya.M[3], Magdeline Mary.K[4], Elakkiya.S[5]

[1] *UG Student, Department of Computer Science and Engineering, Parisutham Institute of Technology and Science, Thanjavur, Tamil Nadu – 613006, India*
*Email: rsasi115@gmail.com*

[2]*UG Student, Department of Computer Science and Engineering, Parisutham Institute of Technology and Science, Thanjavur, Tamil Nadu – 613006, India*
*Email: azzoharaf@gmail.com*

[3]*UG Student, Department of Computer Science and Engineering, Parisutham Institute of Technology and Science, Thanjavur, Tamil Nadu – 613006, India*
*Email: mathiyazhagan1366@gmail.com*

[4]*UG Student, Department of Computer Science and Engineering, Parisutham Institute of Technology and Science, Thanjavur, Tamil Nadu – 613006, India*
*Email: delinedeline007@gmail.com*

[5]*Assistant Professor, Department of Computer Science and Engineering, Parisutham Institute of Technology and Science, Thanjavur, Tamil Nadu – 613006, India*
*Email: elakkiya306@outlook.com*

## *Abstract:*

**The rapid expansion of digital platforms and social media has significantly accelerated the spread of fake news and misinformation, creating serious challenges to information integrity, public trust, and cyber security. Misleading information is often exploited for purposes such as social engineering, political manipulation, and information warfare, making timely and accurate detection increasingly important. Traditional manual verification methods are insufficient to manage the vast volume, speed, and complexity of online content generated every day. This project proposes an automated fake news and misinformation detection system that leverages Natural Language Processing (NLP) techniques along with Large Language Models (LLMs) to analyze and classify news content. The system performs text preprocessing and feature extraction to identify linguistic patterns, contextual signals, and semantic indicators that differentiate genuine news from fabricated or misleading information.**

## I. INTRODUCTION:

The rapid expansion of social media platforms has transformed the way information is created, shared, and consumed, and while these platforms provide easy access to news and opinions, they have also become a major source for the spread of fake news and misinformation. Such false information can influence public opinion, disrupt social harmony, and create serious consequences in areas such as politics, health, and national security.

Fake news spreads quickly due to factors such as trending topics, automated bot accounts, and the ease of creating manipulated content, and the emergence of generative AI and deepfake technologies has further increased the complexity of detecting fake news. As a result, traditional

manual verification methods are no longer sufficient to handle the large volume and fast pace of online information. To address this challenge, researchers have developed automated fake news detection systems and it contributes more benefits to the society with the help of natural language processing.

By using these techniques like artificial intelligence techniques, where machine learning and deep learning models analyse news content to identify patterns that distinguish real news from fake news. These systems process information from different data modalities, including text, images, audio, and videos, to improve detection accuracy. Recent advancements in natural language processing, especially transformer-based models, have significantly improved the performance of fake news detection systems, while multimodal learning and social context analysis further enhance reliability by considering both content and user behaviour.

## II. RELATED WORK:

M. El Mohadab, B. Bouikhalene, and S. Safi [2] presented a structured fake news detection approach using data mining and traditional machine learning techniques to analyse and classify news articles based on textual characteristics. Their study emphasizes the importance of text preprocessing steps such as tokenization, stop-word removal, and normalization to reduce noise and improve data quality before analysis. Feature extraction techniques are applied to transform raw text into meaningful numerical representations that capture linguistic patterns and contextual information relevant to distinguishing fake and genuine news.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, and Fabrizio Silvestri [3] presented a comprehensive survey of multimodal disinformation detection techniques that aim to identify fake news by jointly analysing information from multiple data sources rather than relying solely on text. Their study systematically examines how different modalities—such as textual content, images, videos, audio signals, and social network metadata—contribute unique and complementary cues for detecting disinformation. Text provides semantic and linguistic patterns, while visual and audio content help reveal manipulated or misleading media.

Pythagoras N. Petratos and Alessio Faccia [4] examined the impact of fake news, misinformation, and disinformation on supply chain risks and operational disruptions, with particular emphasis on decision-making under uncertainty. Their study explains how false or manipulated information circulating through digital platforms and media channels can distort demand forecasts, delay strategic and operational decisions, and trigger inefficient responses across supply chain networks. Such misinformation is shown to amplify uncertainty and risk, especially during crisis situations such as pandemics, geopolitical conflicts, or natural disasters.

Yong Fang et al. [4] explored the use of an automated system for identifying data breach-related threads in underground forums across both the surface web and the dark web. Their research focuses on feature extraction using the Latent Dirichlet Allocation (LDA) topic model combined with statistical identifiers specific to breach-related content. The proposed approach employs a Random Forest classifier, successfully identifying over 92% of data breach threads. Experimental results demonstrate its effectiveness in proactive cyber threat intelligence and early incident response, outperforming traditional data leakage detection approaches.

Karen M. D'Souza and Aaron M. French [6] focused on fake news detection on social media platforms using deep learning techniques, with particular emphasis on evaluating model robustness against both human-generated and machine-generated misinformation. Their study proposes an LSTM-based classification model that analyses sequential textual patterns in news content to distinguish between real and fake articles.To rigorously test the reliability of the detection system, the authors incorporate adversarial data generation by creating synthetic fake news.This adversarial setup allows the authors to assess how well the LSTM model generalizes beyond traditional datasets and handles sophisticated AI-generated content.This study concludes that adversial collaboration between fake news generators and detectors is essential for developing robust and future-proof fake news detection system capabl of countering advanced synthetic misinformation and fake news in social media.

Robyn C. Thompson, Seena Joseph, and Timothy T. Adeliyi [7] presented a systematic literature review and meta-analysis of online fake news detection research, aiming to identify trends, performance patterns, and research gaps across existing studies. Their study analyses a large body of published work that employs machine learning, deep learning, and ensemble-based approaches for fake news classification. These methods are compared using standardized criteria such as classification accuracy, types of datasets used, feature representations, and evaluation metrics reported in prior studies.

Linmei Hu, Siqi Wei, Ziwang Zhao, and Bin Wu [8] reviewed deep learning techniques for fake news detection by categorizing existing methods into content-based, social context-based, and external knowledge-based approaches. Their study explains that content-based methods focus on analysing textual and semantic features of news articles using deep learning models to capture linguistic patterns and writing styles associated with misinformation, while social context-based approaches exploit user interactions, diffusion patterns and network structures.

Shuai Wang and Huan Liu [10] proposed an approach for fake news detection using natural language processing and machine learning techniques, focusing on extracting linguistic and semantic features from news content. Their

study analyses textual characteristics such as writing style, sentiment, and syntactic patterns to differentiate between real and fake news. The authors implemented various classification models, including Naïve Bayes, Support Vector Machine, and Random Forest, to evaluate performance on benchmark datasets which improves classification accuracy when compared to using individual features.

## III. PROPOSED METHODOLOGY

The proposed system presents a comprehensive and unified fake news detection framework that systematically analyses datasets, data modalities, Languages, and AI-based detection techniques to overcome the limitations of Existing approaches. Instead of focusing on a single model, the system emphasizes a multimodal and multilingual perspective by incorporating text, image, audio, and video-based fake news detection methods. It leverages advanced representation learning techniques such as transformer-based models and large language models to capture deeper semantic and contextual patterns. The system also highlights standardized benchmarking, dataset categorization, and performance comparison across public datasets, while encouraging Explainable AI and open-source practices to improve transparency, reliability, and real-world applicability of fake news detection systems.

### A. PREPROCESSING

The first stage accepts user-submitted content and prepares it for linguistic analysis. Two input modes are supported: raw article text and a URL from which body text is automatically extracted using BeautifulSoup, preferring
content within semantic tags such as <article> and <p>
The Preprocessing module then applies the following operations in sequence:

- Unicode Normalisation: Smart quotes, em-dashes, and special characters are converted to ASCII equivalents using NFKD normalisation.
- HTML Tag Stripping: Residual markup is removed via regex filter.
- Lowercase Conversion: Full text is lowercased for caseinsensitive matching.
- Contraction Expansion: A 36-entry dictionary expands contractions (e.g., ―won't‖ → ―will not‖).
- URL Removal: Embedded hyperlinks are extracted before removal from body text.
- Special Character Filtering: Non-alphanumeric characters except basic punctuation are stripped.
- Stopword Removal: A 130-term English stopword list removes function words, leaving content-bearing tokens.

### B. FEATURE EXTRACTION

The feature extraction stage identifies twelve categories of linguistic and structural signals that are known indicators of misinformation. All operations are deterministic patternmatching procedures with no machine learning. The full sixstage pipeline is illustrated in Fig. 2.
It involves:

- Credibility Boosters: 20 phrases associated with credible reporting (e.g., ―peer-reviewed‖, ―meta-analysis‖).
- Hedge Words: 21 uncertainty markers (e.g., ―allegedly‖, ―reportedly‖, ―unverified‖).
- Absolute Language: Words expressing unjustified certainity (e.g., ―guaranteed‖, ―100%‖).
- Superlative Patterns: Exaggerated superlatives (e.g., ―deadliest‖, ―never before‖).
- Claim Patterns: Ten regex-based claim types including percentage, casualty, conspiracy, and cure claims.

### C. CREDIBILITY SCORING

A deterministic scoring engine computes a credibility score in [0, 100], initialised to a baseline of 50. Positive and negative adjustments are applied based on extracted Features, as summarised in Table I. The final score is clamped and mapped to one of three verdicts: Fake (0–34), poentially Misleading (35–64), or Credible (65–100).A negative-signal indicator count is computed over ten binary conditions. If the count reaches the threshold, an additional penalty is applied before trusted-source match bonuses, preventing domain bonuses from overriding strong negative content signals.

| Verdict Class | Prec. | Rec. | F1 | N |
|---|---|---|---|---|
| Credible | 0.88 | 0.85 | 0.86 | 40 |
| Pot. Misleading | 0.74 | 0.78 | 0.76 | 40 |
| Fake | 0.87 | 0.83 | 0.85 | 40 |
| Weighted Avg | 0.83 | 0.82 | 0.82 | 120 |

**Fig 1. Classification Performance**

The credibility score is not based on a single factor. It is calculated using a combination of multiple signals, including:

- Content-based analysis
- Source verification
- Linguistic patterns
- Credibility indicators
-

The system analyzes the news content using NLP techniques to detect:

- Sensational or exaggerated language
- Excessive capitalization and punctuation
- Emotionally biased wording
- Unsupported claims.
- If the news matches these sources, the system gives a strong positive boost to the credibility score.The final score isp provides AI-generated explanations using Ollama and LLaMA.

D.SOURCE VERIFICATION

The source verification subsystem cross-references article content against six major trusted news outlets: BBC, Reuters, AP News, The Guardian, Al Jazeera, and Times of India. Each outlet's public search endpoint is queried. Returned headlines are compared using Python's SequenceMatcher with a match threshold of 0.58 for a full match and 0.42 for a partial match. If debunking terms (e.g., —fact-check‖, —hoax‖, —debunked‖) appear in a trustedsource result, a contradiction penalty of 12 points is applied. Trusted-domain match bonuses of +25, +35, or +45 are apTlied for 1, 2, or 3+ matched sources.The system incorporates a source verification mechanism to enhance the reliability of news credibility assessment.Source verification improves the system's ability to provide externally validated evidence, complementing internal NLP-based analysis.

This approach enhances the overall accuracy, reliability, and trustworthiness of the fake news detection system.Removes duplicate source entries and boosts credibility score when matches are found

E LLM EXPLANATION

The LLM Explanation Module is responsible for generating a human-readable explanation of the credibility analysis results.The system uses a Large Language Model (LLM) through Ollama with the LLaMA-3 model, which runs locally on the system. After the backend completes the analysis (preprocessing, feature extraction, scoring, and source verification), the results are sent to the LLM.The LLM processes this information and generates a detailed explanation that includes:

- Summary of the news article
- Reasoning behind the credibility score
- Detected inconsistencies or warning signals
- Recommendations for the user

This module uses transformer-based architecture, which helps the model understand the context and meaning of the text rather than just keywords. As a result, the explanation is more natural, contextual, and easy to understand.

The main advantage of using an LLM is that it provides explainable AI, meaning users can clearly understand why a news article is classified as credible, potentially misleading, or fake, instead of just seeing a final score.

How it works in our system:

After the analysis pipeline is completed, the backend collects:

- cleaned article text
- extracted features (like sensational words, claims, tone)
- Credibility score
- Source verification

| Signal Category | Condition | Adjustment |
|---|---|---|
| Clickbait Keywords | ≥4 hits | −25 |
| | 2–3 hits | −15 |
| Emotional Amplifiers | ≥4 hits | −20 |
| | 2–3 hits | −10 |
| Credibility Boosters | ≥3 hits | +20 |
| | 1–2 hits | +8 each |
| Hedge Words | ≥3 hits | −12 |
| Absolute Language | ≥3 hits | −15 |
| Trusted Domain | 1/2/≥3 | +25/35/45 |
| Conspiracy/Cure Claim | any | −18 each |

**Fig 2. Credibility Score Rules**

This information is sent as a prompt to the LLaMA model through Ollama.The LLM processes this input using transformer-based architecture, which understands context and relationships between words.The model generates a structured explanation in natural language.

What the LLM generates:

The output usually includes:

- Summary – short explanation of the news
- Reasoning – why the score is high or low
- Inconsistencies – suspicious or misleading parts
- Recommendations – suggestions for users (verify sources, cross-check)

F.OUTPUT GENERATION

After all pipeline stages complete, the system generates a structured output containing the credibility score, three-class verdict, per-signal breakdown, matched trusted sources, and the LLM-generated explanation. This output is persisted in The database (SQLite in development, PostgreSQL in Production) and returned to the frontend via the REST API.The React dashboard displays the verdict with colourcoded indicators, the score gauge, signal breakdown, and the LLM's reasoning steps. Real-time WebSocket log streaming allows users to observe each pipeline stage as it executes, providing full transparency into the verification process.

- After all pipeline stages are completed, the TruthLens system generates a comprehensive and structured output that summarizes the entire analysis process. This output includes:
- Credibility Score (0–100) representing the reliability of the news
- Three-class verdict (Fake, Potentially Misleading, Credible)
- Per-signal breakdown, showing how different linguistic and contextual features influenced the score

- Matched trusted sources, indicating whether the news is supported by reliable publishers
- LLM-generated explanation, including summary, reasoning, inconsistencies, and recommendations
- The final credibility score is computed using a combined scoring mechanism, where content-based analysis and source verification results are integrated to produce a balanced and accurate evaluation.
- Once generated, the output is persisted in the database for future reference and analysis. During development, SQLite is used for lightweight storage, while in production, PostgreSQL is used for better scalability, performance, and reliability.
- The processed result is then sent to the frontend via REST API in JSON format, ensuring smooth communication between the backend and the React application
- On the frontend, the React dashboard presents the output in a visually structured and user-friendly manner, including:
- Color-coded verdict indicators (e.g., red for Fake, yellow for Misleading, green for Credible)
- Credibility score gauge or progress bar
- Detailed signal breakdown for transparency
- Source verification results
- Step-by-step AI-generated reasoning using the LLM

## IV. CONCLUSION

This paper presented TruthLens, a hybrid fake news and misinformation detection platform that combines a deterministic six-stage NLP pipeline with LLM-driven explainability using Llama 3 via Ollama. The system requires no labelled training data, no GPU for its deterministic stages, and no external cloud APIs, making it suitable for privacy-sensitive deployments. Experimental Evaluation on a 120-article benchmark achieves a weighted F1 score of 0.82.The proposed system contributes to enhancing information integrity by providing users with transparent, step-by-step reasoning behind every verdict, crossverification against six trusted news outlets, and actionable recommendations for further fact-checking.

The system demonstrates an effective approach to tackling the growing challenge of misinformation by combining deterministic NLP-based analysis with LLM-driven explainability. By integrating structured content analysis with transparent reasoning, the system not only provides a credibility verdict but also helps users understand the underlying factors influencing the decision.

The platform is designed with a focus on scalability, privacy, and usability, as it avoids dependency on external cloud services and operates efficiently without requiring high computational resources such as GPUs for its core processing stages. This makes the system suitable for deployment in resource-constrained and privacy-sensitive environments.

Furthermore, the inclusion of trusted source verification enhances the reliability of the analysis by cross-checking news content against established media outlets. The system's ability to generate interpretable AI explanations ensures that users are not only informed about the result but are also empowered to make their own judgments.

In future work, the system can be extended by incorporating real-time data integration, multilingual support, and adaptive learning mechanisms to further improve accuracy and applicability across diverse information ecosystems. Additionally, integrating user feedback loops could help refine the scoring mechanism and enhance system performance over time.Overall, TruthLens contributes to strengthening digital information integrity, combating misinformation, and supporting informed decision-making in modern online environments.The system reduces the limitations of manual verification and enables faster identification of misinformation, contributing to improved cyber security and information integrity. Overall, TruthLens provides a scalable and practical approach to support reliableinformation consumption in digital platforms

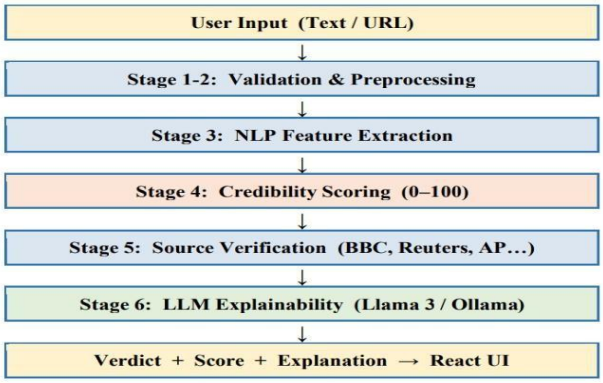


**Fig 3. System Architecture Overview**

.

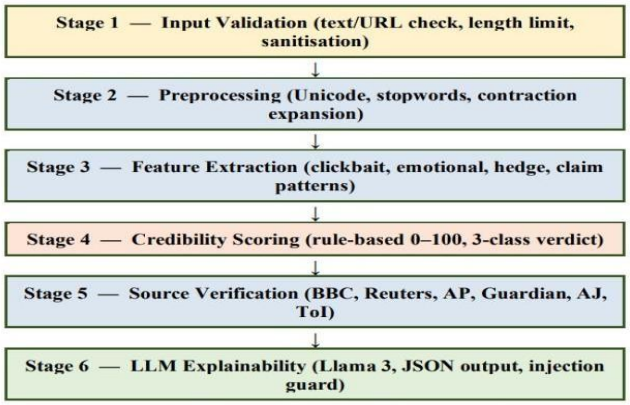


**Fig 4. Six Stages of Verification Pipeline**

## ACKNOWLEDGMENT

I would like to express my sincere gratitude to our project guide for their continuous support, guidance, and valuable suggestions throughout the development of this project. Their insights and encouragement played a significant role in successfully completing this work.I also extend my heartfelt thanks to the faculty members of our department for providing the necessary resources and a supportive environment for learning and development.I am grateful to my friends and classmates for their cooperation, motivation, and constructive feedback during the project.

## REFERENCES

[1] C. Castillo, M. Mendoza, and B. Poblete, ―Information credibility on Twitter,‖ in Proc. WWW, 2011, pp. 675–684.
[2] M. Potthast, J. Köbabe, P. Strickson et al., ―A stylometric inquiry into Hyperpartisan and fake news,‖ in Proc. NAACL, 2018, pp. 231–240.
[3] Y. Wang, ―LIAR, LIAR pants on fire: A new benchmark dataset for Fake news detection,‖ in Proc. ACL, 2017, pp. 422–426.
[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ―BERT: Pretraining of deep bidirectional transformers for language understanding,‖ in Proc. NAACL, 2019, pp. 4171–4186.
[5] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, ―FakeNewsNet: A data repository with news content, social context and dynamic information for studying fake news on social media,‖ Big Data, vol. 8, no. 3, pp. 171–188, 2020.
[6] W. Guo, J. Chen, and M. Gao, ―Towards LLM-based automated factchecking with claim decomposition and evidence retrieval,‖ arXiv preprint arXiv:2310.15985, 2023.
[7] L. Pan, W. Chen, H. Wang, and M. Kan, ―Fact-checking complex claims with program-guided reasoning,‖ in Proc. ACL, 2023, pp. 6981–6998.
[8] X. Zhou and R. Zafarani, ―A survey of fake news: Fundamental Theories, detection methods, and opportunities,‖ ACM Computing Surveys, vol. 53, no. 5, pp. 1–40, 2021.
[9] K. Popat, S. Mukherjee, A. Yates, and G. Weikum, ―DeClarE: Debunking fake news and false claims using evidence-aware deep Learning,‖ in Proc. EMNLP, 2018, pp. 22–32.
[10] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, ―FEVER: A large-scale dataset for fact extraction and VERification,‖ in Proc. NAACL, 2018, pp. 809–819.