MULTILINGUAL NLP AND THREAT DETECTION

A Literature Review on Multilingual Fake News and Scam Detection for Regional Languages Using Natural Language Processing

Neeraja Sujaysinh Sankpal, Janhavi Santosh Redekar, Snehal Kishor Sankpal, Sanyukta Sudhir Patil, A. D. Patil

Department of Computer Science and Engineering, Sanjay Ghodawat Institute, Atigre, Kolhapur, Maharashtra, India

Published November 2026

Abstract

Sharing news, messages and links online now takes only a few seconds. The same speed helps false stories, misleading posts, scam messages and harmful links travel across social networks and chat apps. Researchers have tried to catch such content with machine learning, deep learning, transformer-based models, multimodal techniques and automatic fact-checking. However, much of this effort has been built around English and depends on very large datasets, so it does not carry over easily to Indian regional languages like Hindi and Marathi.

In this paper we look at earlier work on detecting fake news, processing text in many languages, combining text with images, checking facts automatically, explaining model decisions and spotting suspicious web links. We also consider how people actually write in regional languages, including mixed-language messages, words typed in Roman script and the many formats in which content is shared. From these studies we note several weak points: few regional-language resources, results that are hard to interpret, limited support for different input types, little source checking and no single tool that brings threat detection together.

To address these points we put forward RegionalGuard, a framework that examines digital content written in English, Hindi, Marathi and Hinglish. It brings together language detection, message checking, text extraction from screenshots, link inspection, news verification, review of social-media posts and forwarded messages, risk scoring and results written in simple words. The tool is meant mainly for people who use regional languages in their daily communication.

Keywords: fake news detection; regional languages; multilingual NLP; scam detection; explainable AI; URL analysis

1 Introduction

The way people receive information has been reshaped by smartphones, social networking sites, chat applications and online news portals. These channels are fast, but they also allow wrong or harmful material to reach a large audience in very little time. Typical examples are made-up news stories, false statements, edited information, fake government notices, bogus job offers, lottery messages and dangerous links. Many people pass such content on to friends or groups without checking whether it is true, and this makes the spread even wider.

Detecting fake news automatically has become an important research topic within natural language processing and machine learning. The first studies in this area mostly dealt with building benchmark datasets and classifying text. Wang released the LIAR dataset, which holds 12.8K short statements

labeled by hand, and evaluated a hybrid convolutional neural network that drew on both the text and its metadata [1]. Shu et al. treated the problem as a data mining task, studying the content of news items along with the social setting in which they were shared [2].

Subsequent studies broadened this view. Zhou and Zafarani sorted the available methods into four groups based on false knowledge, style of writing, the way content spreads and how credible the source is [3]. D'Ulizia et al. put several fake-news datasets side by side and observed that they vary in subject area, labeling, features and testing conditions [4]. Shah and Ganatra surveyed machine learning and deep learning methods and pointed out unresolved issues related to data, algorithms and evaluation [5].

A large share of these systems were built for English. Indian regional languages have far less data and fewer language tools available, and people frequently blend Hindi or Marathi with English inside the same message. For Hindi, researchers have tried CNN-LSTM models on news text [6], while multilingual work has applied attention-based techniques over a number of languages [7]. Any detector meant for this setting must be able to deal with differences in language use and with text that switches between languages.

False content does not always come as plain text. It often appears in screenshots, pictures, captions and links, which is why multimodal analysis has grown into a separate line of research. Fakeddit is a large multimodal benchmark that includes text, pictures, metadata and detailed category labels [8]. Segura-Bedmar and Alonso-Bartolome reported that using text and images together gives better fake-news classification than using either alone [9].

Tools and data for Indian languages have also expanded. IndicNLPSuite offers text corpora, word embeddings, pretrained language models and test resources for 11 widely spoken Indian languages [10]. In 2025, a study released a fake-news dataset that is both multilingual and multimodal, spanning Hindi, Gujarati, Marathi, Telugu and English, and presented the HEMT-Fake framework alongside it [11].

Apart from false news, people also receive scam messages and harmful links. Researchers have applied machine learning to find phishing URLs [12], and other work has relied on hand-crafted features and deep learning models for the same purpose [13]. Automatic fact-checking adds another layer by picking out statements that should be verified, even when they are written in different languages [14].

Fake-news detection, multilingual NLP, multimodal analysis, URL detection and fact-checking are, for the most part, researched as independent topics. In this review we go through the current methods in each area and point out the space where a single multilingual framework could be useful.

2 Literature review

2.1 Traditional and content-based fake news detection

The earliest work on fake news was mainly about creating trustworthy datasets and understanding how misleading text is written. Wang built the LIAR benchmark from 12.8K short statements taken from PolitiFact and labeled by hand, and proposed a hybrid convolutional neural network that used text features together with metadata [1]. LIAR went on to become a widely used benchmark. Since it is made up of short English statements, though, it cannot be applied as it is to multilingual or regional-language messages.

Shu et al. viewed fake-news detection through the lens of data mining and studied both what the news says and the social surroundings in which it circulates [2]. Their work showed that the way users interact with a post and the way it spreads can reveal clues that the text by itself does not. This is relevant for any system that inspects social-media posts or forwarded messages.

Zhou and Zafarani carried out a survey of detection methods and arranged them according to false knowledge, writing style, spreading behavior and trustworthiness of the source [3]. Their grouping implies that a good detector might have to consider more than one of these aspects together.

D'Ulizia et al. studied the datasets that are usually used to test fake-news detectors [4]. These datasets are not alike in topic, labels, features or test setup, and such differences change the performance figures that papers report. For this reason, a system aimed at Indian regional languages needs data that matches the way its own users write, rather than depending only on English collections.

2.2 Machine learning and deep learning approaches

Shah and Ganatra examined machine learning and deep learning methods together with the datasets, evaluation metrics and open issues linked to them [5]. Their review notes that standard machine learning can give good classification results, while deep learning is better at picking up complicated patterns in text.

In 2025, a study focused on Hindi applied a CNN-LSTM model to a newly collected set of 6,724 Hindi news articles [6]. The work highlights the value of data and models designed for a particular language. Its scope, however, is restricted to Hindi news, and it does not deal with suspicious links, screenshots or scam messages.

Ahuja and Kumar proposed Mul-FaD, a multilingual detector based on attention that was trained using news in English, German and French [7]. The attention layers allowed the model to focus on the most useful parts of the text in each language. Because no Indian regional language was part of the study, its usefulness for Hindi or Marathi has not yet been tested.

2.3 Multimodal fake news detection

More and more false content mixes text with pictures, screenshots and captions. Nakamura et al. presented Fakeddit, a large multimodal benchmark for detailed fake-news detection, which holds text, images, metadata and user comments and allows classification at more than one level [8]. Working on Fakeddit, Segura-Bedmar and Alonso-Bartolome compared models that use only text with models that also use images, and found that the image data improved results [9].

On the other hand, multimodal systems demand more data, more processing power and more complicated designs than systems that look at text alone. A regional-language tool meant for real use must therefore weigh how well it detects threats against these extra costs and against how simple it is to use.

2.4 Multilingual natural language processing

Finding misinformation in Indian languages calls for models that can work with several scripts, word stocks and sentence structures. Kakwani et al. developed IndicNLPSuite, which gathers corpora, embeddings, pretrained language models and test resources for 11 major Indian languages [10].

A 2025 study released a dataset that is multilingual as well as multimodal and covers Hindi, Gujarati, Marathi, Telugu and English. The same study proposed HEMT-Fake, an explainable framework that brings together text, images and relational information [11].

Nandgaonkar and Mane used a hierarchical attention network for English, Hindi and Marathi [15]. They prepared a Marathi dataset and studied attention at the level of words as well as sentences. Their findings favor hierarchical models for languages with few resources, and at the same time show how scarce Marathi data still is.

2.5 Explainability and automated fact-checking

Simply labeling content as fake or real gives users no idea of why it was marked. Shehu et al. developed an explainable fake-news detector that uses particle swarm optimization and mutual information to choose the most informative features, which makes its decisions easier to understand [16].

Jadhav presented HEMT-Fake, an explainable fake-news detection framework for Hindi, Gujarati, Marathi and Telugu, with English also part of the dataset used in the study. The framework merges text, images and relational data and adds explainability methods on top [11].

For fact-checking, a system first has to locate the claims that deserve checking and then collect evidence that supports or contradicts them. Panchendrarajan and Zubiaga surveyed claim detection for single-language, multilingual and cross-lingual fact-checking, and separated the step of identifying a claim from the step of verifying it [14]. Other researchers have turned to large language models to check facts in many languages and to verify claims across languages [17].

2.6 Scam and phishing URL detection

Scam messages frequently carry links that lead to fake or harmful websites. Hasane Ahammad et al. evaluated machine learning methods for spotting phishing URLs and showed that the properties of a URL are useful for recognizing dangerous links [12]. Ferdaws and Ebrahimi Majd looked at feature-based machine learning and deep learning methods, among them CNN and LSTM models [13]. A further IEEE study suggested a web-based machine learning framework that joins URL classification with checks on embedded links and WHOIS domain records [18]. Checking URLs is therefore a helpful extra step when judging whether content is unsafe.

Looking across all of these areas, most of the abilities discussed are available only as separate tools. A digital-safety system that joined them together would allow users to examine doubtful messages, screenshots, links, news claims, social-media posts and forwarded content from a single place.

3 Comparison of existing studies

Table 1. Comparison of existing studies.

Author(s)	Year	Methodology	Advantage	Limitation
Wang [1]	2017	Hybrid CNN using NLP features and metadata	Created the LIAR benchmark and joined text with metadata	Deals mostly with short English statements
Shu et al. [2]	2017	Data mining, NLP and study of social context	Looks at both news content and social signals	Gives little attention to regional languages

Author(s)	Year	Methodology	Advantage	Limitation
Zhou and Zafarani [3]	2020	Review of fake-news detection methods	Covers content, spread patterns and source trust	A review, not a working detection system
D'Ulizia et al. [4]	2021	Review and analysis of datasets	Points out how test datasets differ	No full detection system is proposed
Shah and Ganatra [5]	2022	Systematic review of ML, DL and NLP work	Sums up algorithms, datasets and open issues	Little coverage of Indian regional languages
Gupta et al. [6]	2025	CNN-LSTM model for detecting Hindi fake news	Makes use of a dataset built for Hindi	Restricted mostly to Hindi news
Ahuja and Kumar [7]	2023	Multilingual detection based on attention	Handles fake news in more than one language	Indian regional languages are not its focus
Nakamura et al. [8]	2020	Detailed multimodal detection	Offers a large multimodal benchmark	Heavy to compute and not aimed at regional languages
Segura-Bedmar and Alonso-Bartolome [9]	2022	Text-only and multimodal comparison	Uses both text and image information	Needs multimodal data and extra processing
Kakwani et al. [10]	2020	IndicNLP and IndicBERT	Works with many Indian languages, Hindi and Marathi among them	Not a full fake-news detection system
Patil et al.	2024	Building a multilingual dataset	Covers Hindi, Marathi, Gujarati and Telugu	Offers a dataset but no combined tool for users
Bhosle et al.	2025/2026	Multilingual ML classification with source checking	Handles four Indian languages and cross-checks sources	Concentrates mainly on classifying fake news
Jadhav [11]	2025	Explainable detection across languages and modalities	Works across languages and modalities with explanations	Needs a complex design and bigger datasets
Hasane Ahammad et al. [12]	2022	Machine learning to detect phishing URLs	Targets the identification of harmful URLs	Covers mostly phishing URLs only
Ferdaws and Ebrahimi Majd [13]	2024	ML and DL with CNN and LSTM models	Shows how well URLs can be detected	Does not include analysis of regional-language messages
Panchendrarajan and Zubiaga [14]	2024	Automatic fact-checking and claim detection	Reviews claim detection in one, many and cross languages	Concentrates mostly on detecting claims
Nandgaonkar and Mane [15]	2026	Hierarchical attention network model	Covers English, Hindi and Marathi	Marathi data is still scarce
Shehu et al. [16]	2026	Explainable feature selection with PSO and mutual information	Better feature choice and easier interpretation	Not a combined regional digital-safety tool

Author(s)	Year	Methodology	Advantage	Limitation
Singhal et al. [17]	2024	Multilingual fact-checking with LLMs	Verifies facts in several languages	Depends on reliable evidence and language resources
Choutapally and Das [18]	2025	URL and domain analysis using ML	Adds domain information to URL classification	Concentrates mostly on phishing and URL threats

4 Research gap

The bulk of fake-news research so far has been carried out on English data. Newer studies have begun to look at multilingual and Indian-language content, yet resources for Marathi and Hindi are still limited. Text in regional languages also comes with its own difficulties, such as varied spellings, mixing of languages, words written in another script and casual writing.

Current systems tend to deal with only one kind of content, whether that is news articles, social-media posts, images or links. In everyday use, people come across all of these together. A forwarded message can carry a link, a screenshot can show a doubtful offer, and a social-media post can state a claim that nobody has checked. None of the systems we reviewed handles all of these formats in one place.

Deep learning and transformer models perform well at spotting misinformation, but they require big datasets and a lot of computing power. This makes them difficult to run in light applications, in schools or on systems with modest hardware.

People also want to understand the reason a message was marked as risky. A useful system should name the exact warning signs, for example pressure to act quickly, promises of prizes or lottery wins, requests for private details, suspicious words, unsafe links or a mismatch with reliable sources.

Fact-checking and source verification are generally researched separately from threat detection at the message level. Bringing news verification and source comparison together with content analysis would give users a more complete view of what they are looking at.

As a result, scam detection, fake-news identification, screenshot analysis, link checking, social-media analysis and forwarded-message analysis are still treated as separate problems. This shows the need for a light, multilingual and explainable framework that brings them onto one platform for users of Indian regional languages. RegionalGuard has been designed with this aim, and it presents its findings as risk levels along with the reasons behind them.

5 Proposed work: RegionalGuard

RegionalGuard is a digital-safety framework built from separate modules that examines online content which may be suspicious or misleading. It works with English, Hindi, Marathi and Hinglish and gives users clear results based on risk.

5.1 Input sources

A user can type in a message, upload a picture or screenshot, paste a URL, submit a social-media post, check a forwarded message or enter a news claim to be verified. Together these options cover situations that a detector working only on text would miss.

5.2 Language identification

As a first step, the system works out which language the content is written in: English, Hindi, Marathi or Hinglish. This is important for regional-language content, because script, vocabulary, spelling and mixed-language phrases can differ a great deal.

5.3 Threat detection and risk assessment

Next, the content is examined for suspicious features. At present the system relies on rule-based detection instead of a trained machine learning classifier. These rules search for signs of fake job offers, lottery and prize scams, demands for money or login details, false claims linked to the government and other doubtful information.

The result of the analysis includes a threat category, a risk level, a score, a confidence value and an explanation. Content is placed in one of three risk levels, namely Safe, Caution and High Risk, which gives the user more useful information than a simple real or fake label.

5.4 Screenshot-based analysis

Doubtful content is often shared as a picture, so its text cannot be copied directly. An OCR module reads the text from the uploaded image and sends it on to the language detection and threat analysis steps. In this way, messages received as images are checked in the same manner as typed messages.

5.5 URL assessment

The URL module inspects the web addresses that users submit for features commonly found in unsafe links. This lets the system judge both the message itself and any link included in it.

5.6 News claim verification

This module compares claims with the news sources that are available. A claim is not marked as false just because no support is found for it. Instead, the module tells the user whether sources exist and whether they agree, so that the user can decide if more checking is needed before believing or sharing the claim.

5.7 Social-media and forwarded content

Social networks and messaging apps are among the main routes through which misleading content travels. For this reason, RegionalGuard includes separate modules for social-media posts and for forwarded messages. Each module deals with its own input format while using the common language detection and risk assessment steps.

5.8 Explainable output

The results page displays the detected language, the threat category, the risk level, the score, the confidence value and the reason behind the decision, written in a way that ordinary users can understand. A “Think Before You Click” message reminds users to look carefully at suspicious content before they act on it.

5.9 Scan history and data management

Every analysis is saved in an SQLite database, including the scan type, detected language, category, risk level, score, confidence, the content that was checked and the time of the scan. These saved records are used for the history feature, dashboard figures and analysis views.

5.10 Overall system workflow

The user first submits one of the supported kinds of input. The system then detects the language and carries out any needed text extraction or preprocessing. After that, the suitable analysis module is run, the risk is assessed and an explanation is prepared, and finally the scan is saved for history and analytics. The complete flow is shown in Figure 1.

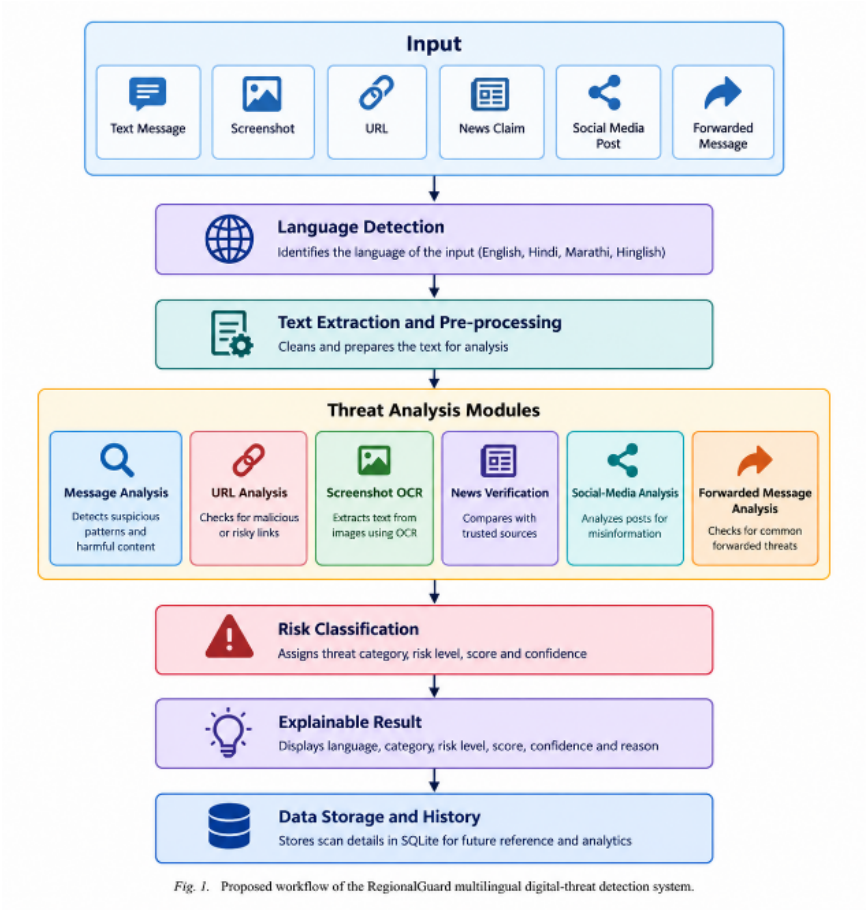


Figure 1. Proposed workflow of the RegionalGuard multilingual digital-threat detection system.

The analysis module that runs depends on the type of input, while every module uses the same language processing, risk assessment and explanation steps. This design keeps the system lightweight even though it handles several languages and input formats.

6 Conclusion

Research on fake-news detection has shifted from linguistic features and classical machine learning toward deep learning, multilingual NLP, multimodal analysis, explainable methods and automatic fact-checking. Benchmarks like LIAR and Fakeddit have made evaluation easier, and IndicNLPSuite together with newer Indian-language datasets has made work on regional languages more practical.

The studies reviewed here show that misleading content in Hindi and Marathi can be detected. Even so, real-world use is still limited by scarce regional-language data, the difficulty of moving models from one language to another, the need for results that people can understand and restrictions on hardware. Work on multimodal detection shows that text and images provide signals that support each other, and work on phishing shows that links should be checked together with the message that carries them.

The key gap is that there are few lightweight, combined tools that bring together regional-language content analysis, OCR on screenshots, link checking, news verification, social-media analysis and forwarded-message analysis. RegionalGuard is put forward to close this gap and supports English, Hindi, Marathi and Hinglish.

In its present form, the system offers rule-based threat detection, language detection, risk assessment, source verification and explanations in simple language. It is intended as a first-level screening and awareness tool and does not replace professional fact-checkers. Its purpose is to help users notice doubtful content and check it before they trust it, reply to it or forward it.

7 Future scope

Building a bigger and carefully checked dataset of regional-language messages and false content would make it possible to train and test machine learning and deep learning models. Multilingual transformer models made for Indian languages could then help the system better grasp context and meaning.

More Indian languages could be added to the framework, along with improved handling of mixed-language and transliterated text. Multimodal methods could also be used to study screenshots, text and metadata together so that more complex misinformation can be handled.

The explanations could be made richer by highlighting the words or phrases that most affected the result, the suspicious features of a link, signs of how reliable a source is and other supporting evidence. Connecting the system to trusted fact-checking services, verified databases and a wider set of news sources would also strengthen claim verification.

Further additions could include analysis that works offline or on weak networks, voice-based use in Marathi and Hindi, more detailed checks on domains and links, and GIS-based regional information. Optional community reporting and IoT-based add-ons could gather digital-safety information that is relevant to a particular area. Taken together, these improvements would make RegionalGuard more helpful across different languages, regions and network conditions, particularly for people in rural and semi-urban areas.

References

[1] Wang WY. 'Liar, liar pants on fire'*: A new benchmark dataset for fake news detection*. In: Proc. 55th Annu. Meeting Assoc. Comput. Linguistics; 2017. p. 422–426. https://doi.org/10.18653/v1/P17-2067

[2] Shu K, Sliva A, Wang S, Tang J, Liu H. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explor Newsl*. 2017;19(1):22–36. https://doi.org/10.1145/3137597.3137600

[3] Zhou X, Zafarani R. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput Surv*. 2020;53(5):109. https://doi.org/10.1145/3395046

[4] D'Ulizia A, Caschera MC, Ferri F, Grifoni P. Fake news detection: A survey of evaluation datasets. *PeerJ Comput Sci*. 2021;7. https://doi.org/10.7717/peerj-cs.518

[5] Shah MN, Ganatra A. A systematic literature review and existing challenges toward fake news detection models. *Soc Netw Anal Min*. 2022;12(1):168. https://doi.org/10.1007/s13278-022-00995-5

[6] Gupta RK, Sharma V, Pateriya RK, Dehalwar V, Gupta P. Fake news detection in Indian languages: A case study with Hindi using CNN-LSTM. *Procedia Comput Sci*. 2025;259:150–160. https://doi.org/10.1016/j.procs.2025.03.316

[7] Ahuja N, Kumar S. Mul-FaD: Attention based detection of multiLingual fake news. *J Ambient Intell Humaniz Comput*. 2023;14(3):2481–2491. https://doi.org/10.1007/s12652-022-04499-0

[8] Nakamura K, Levy S, Wang WY. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In: *Proc*. 12th Lang. Resour. Eval. Conf. (LREC); 2020. p. 6149–6157.

[9] Segura-Bedmar I, Alonso-Bartolome S. Multimodal fake news detection. *Information*. 2022;13(6):284. https://doi.org/10.3390/info13060284

[10] Kakwani D, Kunchukuttan A, Golla S, C. GN, Bhattacharyya A, Khapra MM, et al. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*; 2020. p. 4948–4961. https://doi.org/10.18653/v1/2020.findings-emnlp.445

[11] Jadhav R. Explainable multilingual and multimodal fake-news detection: Toward robust and trustworthy AI for combating misinformation. *Front Artif Intell*. 2025;8:1690616. https://doi.org/10.3389/frai.2025.1690616

[12] Hasane Ahammad SK, Kale SD, Upadhye GD, Pande SD, Venkatesh Babu E, Dhumane AV, et al. Phishing URL detection using machine learning methods. *Adv Eng Softw*. 2022;173:103288. https://doi.org/10.1016/j.advengsoft.2022.103288

[13] Ferdaws R, Ebrahimi Majd N. Phishing URL detection using machine learning and deep learning. In: *Proc*. IEEE World AI IoT Congr. (AIIoT); 2024. p. 485–490. https://doi.org/10.1109/AIIoT61789.2024.10579005

[14] Panchendrarajan R, Zubiaga A. Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. *Nat Lang Process J*. 2024;7:100066. https://doi.org/10.1016/j.nlp.2024.100066

[15] Nandgaonkar S, Mane S. Hierarchical attention network for multilingual fake news detection. *Discover Artif Intell*. 2026;6:1417. https://doi.org/10.1007/s44163-026-02425-3

[16] Shehu HA, Teal PD, Frontin J, Morssy A, Krewel M, Kleijn WB. An explainable approach to fake news detection using PSO and mutual information. *Front Artif Intell*. 2026;9:1901946. https://doi.org/10.3389/frai.2026.1901946

[17] Singhal A, Law T, Kassner C, Gupta A, Duan E, Damle A, et al. Multilingual fact-checking using LLMs. In: *Proc*. Third Workshop on NLP for Positive Impact; 2024. p. 13–31. https://doi.org/10.18653/v1/2024.nlp4pi-1.2

[18] Choutapally R, Das T. Web-based machine learning framework for phishing URL detection and analysis. In: *Proc*. IEEE Conf. Artif. Intell. (CAI); 2025. p. 970–975. https://doi.org/10.1109/CAI64502.2025.00170