Journal of Intelligent Information Systems
https://doi.org/10.1007/s10844-026-01042-x

RESEARCH

Multimodal misinformation detection across diverse languages using RAG and LLMs

Sheetal Harris[1] · **Vinh Thong Ta**[2,3] · **Marcello Trovati**[4] · **Ghada Nakhla**[1] · **Faiza Latif**[5] · **Ioannis Korkontzelos**[1]

Received: 30 October 2025 / Revised: 15 March 2026 / Accepted: 20 March 2026

Abstract
The rapid spread of multimodal fake news (FN) on Online Social Networks (OSNs) threatens digital information ecosystems, particularly in low-resource languages. Existing multimodal fake news detection (FND) methods are largely limited to high-resource settings, restricting their global applicability. We propose an M&M-RAG, a Multilingual & Multimodal Retrieval-Augmented Generation framework, that leverages Large Vision-Language Models (LVLMs) and Large Language Models (LLMs) to verify news claims across English, Chinese and Urdu. M&M-RAG integrates real-time multilingual evidence retrieval, language-aware prompting, and cross-modal reasoning for fact verification. We further propose Multi-Ax-to-Grind Urdu, the first large-scale, multi-domain multimodal benchmark for FND in Urdu. Experiments on typologically diverse monolingual multimodal datasets demonstrate that M&M-RAG achieves state-of-the-art (SOTA) performance, with 94.6% accuracy and 94.2% F1 score, surpassing models such as SpotFake, MPFN, MMCFND, and Semi-FND. The proposed framework remains robust in zero-shot and cross-lingual scenarios under frozen-model inference without task-specific fine-tuning. The results underscore the scalability and interpretability of LVLM-based approaches for combating multimodal misinformation, particularly in under-represented and typologically diverse languages.

Keywords Multimodal multilingual fake news detection · Large vision-language models (LVLMs) · Retrieval-augmented generation (RAG) · NLP

1 Introduction

Online Social Networks (OSNs) have not only transformed news dissemination, but also facilitated the rapid spread of fake news (FN). Modern FN extends beyond textual misinformation to include manipulated images, memes, and captioned visuals that amplify credibility and engagement. The most significant challenge is targeted dissemination of multimodal FN in various regions and languages. Recent advancements in Large language models (LLMs), such as GPT (Achiam et al., 2023), have made it increasingly difficult to

Extended author information available on the last page of the article

distinguish between human- and machine-generated disinformation (Goldstein et al., 2023). Consequently, the development of robust, automated Fake News Detection (FND) systems is crucial (Harris et al., 2025). Most existing FND research has focused on monolingual and unimodal contexts in rich-resource languages (Zhang et al., 2023; Sormeily et al., 2024), which are insufficient for addressing multimodal, cross-lingual misinformation (Caramancion, 2023; Tufchi et al., 2023). Studies indicate that OSN users are more susceptible to visually enriched and clickbait-style news (Albalawi et al., 2023), underscoring the need for multimodal approaches in low-resource languages (Thaokar et al., 2022; Harris et al., 2023). Moreover, these approaches are neither comprehensive nor generalisable across languages (Jing et al., 2023).

Existing multimodal FND models in high-resource languages (e.g., English, Chinese) rely heavily on handcrafted or post-hoc feature fusion (Suryavardan et al., 2023; Ying et al., 2023). Such methods often overlook semantic dependencies between modalities and suffer from noisy feature interactions (Singh et al., 2023; Wang et al., 2022). Although multimodal variational encoders (Khattar et al., 2019) and deep fusion networks, such as MCAN (Wu et al., 2021) and MFIR (Wu et al., 2023), have improved performance, they remain dataset- or domain-dependent (Song et al., 2021), non-scalable, and language- or event-specific (Jin et al., 2017; Wang et al., 2018). Furthermore, most frameworks neglect evidence retrieval, which is critical for transparent and explainable FND. Recent studies (2024–2025) highlight persistent gaps in scalability, zero-shot generalisation, and low-resource language coverage (Tahmasebi et al., 2024; LekshmiAmmal and Madasamy, 2025; Bansal et al., 2024; Turaga and Namin, 2024; Harris et al., 2025). These limitations often stem from the lack of multilingual multimodal datasets (Harris et al., 2024) and over-reliance on translation-based methods, which fail to capture cultural and linguistic nuances (Bender et al., 2021). This work addresses these limitations by introducing a native-script multimodal Urdu dataset and a unified multilingual FND framework.

To this end, we propose M&M-RAG (Multilingual & Multimodal Retrieval-Augmented Generation), which integrates real-time evidence retrieval, language-aware prompting, and vision-language inference through Large Vision-Language Models (LVLMs) and LLMs. Figure 1 illustrates an example where M&M-RAG analyses an Urdu news claim with a viral image, retrieves cross-lingual evidence, and produces an evidence-based explanation. We evaluate our framework on three typologically diverse multimodal datasets: Weibo (Chinese), Multi-Ax-to-Grind (Urdu), and Twitter MediaEval (English). M&M-RAG achieves state-of-the-art results without language-specific fine-tuning, demonstrating robust zero-shot and cross-lingual generalisation. The key contributions of this work are summarised as follows:

1. **Novel Multimodal Dataset in Urdu**: We curate Multi-Ax-to-Grind[1], the first large-scale multimodal fake news dataset (with 21,715 image-text pairs) for a low-resource language, addressing a major resource-availability gap in multilingual FND research.
2. **Unified Multilingual and Multimodal FND Framework**: We propose M&M-RAG, a novel LVLM-based system that jointly performs evidence retrieval, language-specific prompting, and multimodal reasoning, generating evidence-grounded justifications in Urdu, Chinese, and English.

[1] The Multi-Ax-to-Grind Dataset and the associated M&M-RAG Code are freely available under CC by 4.0 license at: https://figshare.com/s/62b9bbda2464d2059eeb.

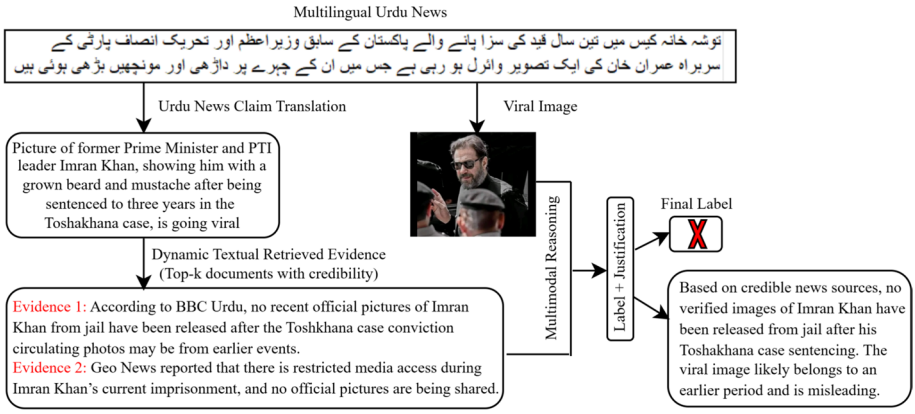


Fig. 1 An example showing the M&M-RAG model that analyses an Urdu news claim (translated into English) with a viral image using dynamic evidence retrieval and Qwen2-VL, predicting it as fake and generating an evidence-based explanation

3. **Zero-shot and Cross-lingual Generalisation**: M&M-RAG demonstrates strong zero-shot performance across distinct languages, marking the first evidence-aware multimodal FND system validated in cross-lingual settings with scalability and linguistic generalisation.

This paper is organised as follows: Section 2 reviews multimodal FND methods in different languages. Sections 3 and 4 detail the contributions of this paper. Section 3 presents the first Urdu multimodal dataset, Multi-Ax-to-Grind Urdu, along with publicly available datasets in English and Chinese. Section 4 entails the proposed M&M-RAG framework for FND. Section 5 presents experimental results and ablation studies, and Section 6 concludes the paper with key findings and future directions.

2 Related work

This section reviews traditional and recent fake news detection (FND) approaches, highlighting their limits in multilingual and multimodal contexts and their focus on high-resource languages. Table 1 provides a summary of reviewed studies.

2.1 Traditional and multimodal FND in monolingual and low-resource languages

The existing FND research focused mainly on monolingual, feature-based approaches that emphasised feature extraction and fusion but paid little attention to cross-lingual generalisation. The importance of unimodal and multimodal coherence is well recognised (Harris et al., 2024). Many studies used complex attention-based fusion mechanisms (Wu et al., 2023), which are effective but lack interpretability and adaptability in multilingual settings.

Various fusion strategies have been explored in multimodal FND. SpotFake (Singhal et al., 2019) employed an early fusion using BERT for textual features and VGG-19 for visual input in Chinese and English datasets, whereas SpotFake+ (Singhal et al., 2020) used

Table 1 Comparison with the existing monolingual and multilingual multimodal studies

Ref & Year	Dataset Languages	Models & Techniques	Contributions	Limitations
Harris et al. (2025)	English, Chinese	CNN, VGG-19	Event-invariant multimodal learning	Lacks social context; no multilingual handling
Hoes et al. (2023)	English, Chinese	Bi-LSTM, VGG-19	Cross-modal feature learning without metadata	Frozen visual encoders; no multilingual handling
Kaliyar et al. (2021)	English, Chinese	BERT, VGG-19	Lightweight framework without extra tasks	Limited fusion techniques explored
Ling et al. (2025)	English	XLNet, VGG-19	Transfer learning with pre-trained models	Fusion strategies not analysed; likely frozen models
Harris et al. (2023)	Chinese, English	CARN, MCN	New dataset; noise-aware multimodal fusion	Dataset-specific; limited language scaling
Kalraa et al. (2021)	English	CNN, BiLSTM, BERT	Multi-class FND with CNN; outperformed unimodal baselines	Early fusion limited to rich-resource languages
Cohen (1960)	English, Chinese	BERT, Swin-T, VGG-19, MLP-Mixer	Progressive fine-grained multimodal fusion	Limited language scope; lacks zero-/few-shot capabilities
Lina et al. (2020)	English	DeBERTa, SwinV2 + ensemble	Unified ensemble for multi-type fusion	No multilingual evaluation; manual ensemble tuning
Brown et al. (2020)	Arabic	MARBERTv2, VGG-19, ResNet-50	First Arabic multimodal dataset	Small imbalanced dataset; weak visual–text correlation
Gravier et al. (2016)	English, Chinese	Ensemble of BERT, ELECTRA, NASNet	Efficient decision-level ensemble & feature fusion	No attention-based fusion; no multilingual generalisation
Li et al. (2023)	7 Indic languages	MuRIL, BLIP-2, FLAVA, NASNet	Large multilingual dataset for similarity-based fusion	English translation required; loses language-specific context
Liu et al. (2023)	Tamil	LLMs, mBERT, XLM-RoBERTa, ViT, DeiT	Curated dataset; LLM-based visual explanation with justifications	Relatively small dataset; no multilingual generalisation
Ours	**English, Urdu, Chinese**	**LVLM-based Language-aware Reasoning**	**Curated multimodal Urdu dataset; Multilingual FND with Evidence Retrieval**	**Claim Translation into English due to the limited model's native support for Urdu**

XLNet instead of BERT for textual encoding in English. SEMI-FND model (Singh et al., 2023) proposed a stacked ensemble using lightweight pre-trained models (BERT, ELECTRA, and NASNet). CARMN (Song et al., 2021) applied CARN with a multi-channel CNN to enhance cross-modal interaction for Chinese and English datasets. Likewise, MPFN (Jing et al., 2023) presented a progressive multi-tier fusion to capture modality-specific representations for Chinese and English. However, the method requires training on each dataset and lacks zero-shot generalisation capabilities, interpretability, and external knowledge as suggested in a recent study (Li et al., 2025). Conversely, Albalawi et al. (2023) explored variance-aware feature extraction and fusion strategies through early and late fusion. The study showed that larger multilingual datasets can enhance performance and generalisation. While these models significantly improved classification performance, they often lack scalability and generalisation across languages or modalities. These fusion-focused models induced multimodal boundaries but remained limited in multilingual adaptability and required extensive task-specific tuning.

In terms of architecture, MVAE (Khattar et al., 2019) introduced variational autoencoders for joint multimodal encoding but relied heavily on dataset-specific training and lacked multilingual adaptability. EANN (Wang et al., 2018) addressed event-specific overfitting by learning event-invariant features through adversarial training on English and Chinese datasets. A multimodal CNN, proposed by Segura-Bedmar and Alonso-Bartolome (2022), demonstrated strong performance in six-class classification for English news, evaluating BERT, BiLSTM, and CNN for text and outperforming unimodal models. The multimodal CNN outperformed unimodal models, and the study suggested that pre-trained vision models and late fusion strategies are better for modality alignment, as explained in Tufchi et al. (2023). The research study (Du et al., 2023) integrated DeBERTa-large with Swin Transformer v2 and handcrafted metadata through adapter-based learning and a unified ensemble to handle modality-type combinations for FND in English news. These models reflect diverse multimodal architectures but are mostly supervised and lack multilingual generalisation capabilities. A recent study proposed MUGCL (Nie and Zeng, 2025), which integrated text, images, and social propagation networks into a single contrastive learning framework. The model captured the multimodal news signals by building dual propagation graphs. Additionally, the uncertainty-aware contrastive learning module identified inconsistencies between modalities and filtered out deceptive signals common on OSNs. Uppada et al. (2023) analysed and jointly modelled three complementary modalities. The authors demonstrated that combining heterogeneous visual and textual signals exhibited improved robustness in detecting misleading imagery and clickbait-style captions, underscoring the importance of multimodal reasoning for reliable FND in contemporary OSN environments.

Language-based analysis has been limited to English and Chinese, as suggested in Jing et al. (2023); Singh et al. (2023); Khattar et al. (2019); Wang et al. (2018); Song et al. (2021), which exhibited strong performance but lacked evaluation in multilingual settings. A step toward low-resource contexts was taken in Bansal et al. (2024), where a multimodal dataset in seven Indic languages was introduced. However, due to limitations in FLAVA, all content was translated into English, reducing linguistic context. A recent Tamil-based framework (LekshmiAmmal and Madasamy, 2025) used LLM-generated image descriptions and vision transformers for image-text alignment and reasoning, with the potential for cross-lingual expansion. However, translating datasets into English caused a loss of linguistic nuance and cultural context.

2.2 LLMs-based unimodal and multimodal FND and RAG

NLP-based methods and PLMs, such as BERT and XLM-RoBERTa, have substantially improved text-based FND, particularly for low-resource languages (Harris et al., 2025). The transition from encoder-based PLMs to generative LLMs (e.g., GPT, LLaMA) reduced task-specific fine-tuning requirements, enabling broader linguistic generalisation (Touvron et al., 2023; Brown et al., 2020; Kaliyar et al., 2021). BERT (Devlin et al., 2019) and multilingual PLMs, such as mBERT and XLM-RoBERTa trained on multilingual corpora (Xue et al., 2020), are widely used for NLP-related tasks in low-resource languages, offering better prediction performance. For Urdu, CharCNN-RoBERTa (Lina et al., 2020) achieved strong unimodal results, confirming transformers' superiority over conventional CNNs in low-resource settings. Among multilingual transformers, RoBERTa outperformed ALBERT, XLM-RoBERTa, mBERT, and ensemble methods (Kaliyar et al., 2021; Kalraa et al., 2021).

Several studies (Xu et al., 2024; Yao et al., 2023; Hoes et al., 2023; Cheung and Lam, 2023) have shown that LLMs, such as GPT (Achiam et al., 2023), LLaMA (Touvron et al., 2023), and Mistral-7B (Jiang et al., 2023), can perform multilingual FND through instruction- or evidence-tuning, although they often remain unimodal and lack transparency. Similarly, Wang et al. (2022) investigated LLM-enhanced textual FND prediction performance. The study demonstrated that GPT-3.5 showed poor performance compared to task-specific Small Language Models (SLMs). However, analysing the external evidence with reasonable justifications enhanced SLMs' performance in news comprehension. A style-agnostic FND mechanism, proposed in Cheung and Lam (2023), addressed LLM-driven style attacks through three components. Despite its robustness, the model is limited to text-based rich-resource languages and lacks multimodal or multilingual capabilities. The DKFND framework (Liu et al., 2024) used LLMs for few-shot FND on English datasets by combining evidence to enhance reasoning and prediction accuracy.

With the rise of LVLMs, FND has evolved from text-based to multimodal contexts. Models, such as LLaVA (Liu et al., 2023) and BLIP-2 (Li et al., 2023), effectively extract and interpret visual features, while benchmarks have evaluated LVLMs on diverse reasoning tasks, including chart fact-checking (Akhtar et al., 2023), OCR (Xu et al., 2024) and image captioning (Cui et al., 2024). Fine-tuning GPT-4 with CNN-based visual encoders and external verification modules (Ramya et al., 2024) improved multimodal FND accuracy. Other frameworks, such as prompt-based zero-shot models (Tahmasebi et al., 2024) and FND-LLM (Wang et al., 2024), showed strong generalisation across English and Chinese through tampering detection, co-attention fusion, and rationale generation. Collectively, these studies confirm the feasibility of LLM- and LVLM-based multimodal classification. Building on this progress, our method introduces structured, language-aware prompts that combine the claim, retrieved evidence, and image context to guide LVLM reasoning in multilingual settings.

Retrieval-Augmented Generation (RAG) frameworks enhance LLMs by grounding predictions in external evidence, improving reliability and factuality (Ling et al., 2025). Recent works demonstrated that reliability in RAG-based systems can also be improved by enhancing the retrieval stage itself, for example, through multi-stage retrieval and re-ranking, hybrid dense–sparse retrieval, or retriever adaptation using hard negatives, as explored in frameworks, such as AT-RAFT (Ren et al., 2025) and HyPA-RAG (Kalra et al., 2024). Such approaches aim to reduce retrieval noise and prevent topically related but irrelevant evidence from entering the reasoning pipeline. While effective in text-based and domain-specific question-answering settings, these methods primarily focus on what evidence is retrieved, rather than how retrieved evidence is used during multimodal verification. In contrast, our work focuses on evidence-grounded multimodal reasoning, where reliability depends not only on retrieval relevance but also on source-credibility constraints and cross-modal consistency among the claim, image, and retrieved evidence during LVLM inference. Consistent with this design choice, we treat retrieval as a lightweight candidate selection mechanism and delegate reliability to provenance constraints and multimodal reasoning. Huang et al. (2026) presented FALG, which integrated LLMs with dynamic retrieval and a multi-path co-attention mechanism. The proposed framework demonstrated improved interpretability by explicitly linking factual evidence to classification decisions. Empiri-

cal results across both English and Chinese datasets showed stronger cross-lingual generalisation and heightened sensitivity to subtle forms of misinformation. The MOCHENG framework (Yao et al., 2023) combined retrieval, claim verification, and explanation for multimodal news but did not implement a true RAG architecture and struggled with complex evidence selection. Similarly, Nezafat and Samet (2024) showed that RAG improves LLM generalisation, but their proposed approach remained computationally intensive and lacked real-time interpretability. Despite growing interest in FND, there is still no unified multilingual and multimodal RAG-based approach, largely due to the absence of large-scale labelled datasets spanning multiple languages and modalities (Liu et al., 2023).

To address these, we propose M&M-RAG, a novel framework for multilingual FND that enables dynamic evidence retrieval across three linguistically and culturally distinct languages. It includes dynamic retrieval, structured prompting, and LVLM reasoning in a zero-shot architecture. To our knowledge, this is the first LVLM-based framework for multilingual, multimodal FND that dynamically integrates external evidence and targets low-resource languages.

2.3 Efficiency-oriented model compression

Recent works explored improving the efficiency of deep vision and multimodal models through low-bit quantization and binarization. Methods, such as QuantSR (Qin et al., 2023d) and BiMatting (Qin et al., 2023b), demonstrated that aggressive parameter quantization can significantly reduce computational and memory overhead while maintaining accuracy in vision-centric tasks, such as image super-resolution and video matting. BiBench (Qin et al., 2023c) further provided a comprehensive benchmark analysing the performance trade-offs and limitations of binarized networks across architectures and tasks, while data-free quantization approaches, such as Diverse Sample Generation (Qin et al., 2023a), investigated efficiency under limited or unavailable training data. These approaches are complementary to our work but operate at a different level of abstraction. These works primarily targeted model compression and deployment efficiency, rather than evidence retrieval, multimodal reasoning, or verification reliability. In this work, we prioritise correctness, interpretability, and multilingual generalisation under frozen-model inference. Integrating retrieval-augmented multimodal verification with efficiency-oriented model compression techniques remains an important direction for future research, particularly for edge or resource-constrained deployments.

3 Benchmarking M&M-RAG on multi-ax-to-grind and public multilingual datasets

3.1 Multi-ax-to-grind dataset curation process

Urdu FND is in its nascent stages, with limited resources. Urdu FN classification is challenging due to the absence of suitable benchmark corpora. There is no multimodal dataset in the Urdu language to date. We have curated the first large-scale, multi-domain, and multimodal labelled dataset, Multi-Ax-to-Grind Urdu. The dataset comprises 21,715 multimodal

news headlines from 15 domains (2014 - 2023). The multi-domain news increases the lexical diversity of the curated dataset; therefore, we included news from different regions.

Fact-checked FN was accumulated from authentic Urdu and Hindi websites, i.e., 75% of the total fake news, 8,259 samples. The unavailability of fact-checked FN was the major challenge faced during data collection. Therefore, we have gathered English fact-checked news (only 25% of the total fake news, i.e., 2,753 samples) from the AFP[2] repository and translated it into Urdu. The machine-translated FN was added to maintain a balanced sample size, which improves the automated FND process. While the translated FN is from English, we acknowledge that machine translation may introduce stylistic or cultural artefacts that are not present in native Urdu. We mitigate this by employing native Urdu journalists for review. In future work, we will focus on sourcing native FN content to improve contextual integrity.

A team of professional journalists, each with over five years of experience in news editing and fact-checking, assisted in the quality assurance process of the dataset. They verified semantic accuracy by comparing translated and original English news, refining grammar, tone, and culturally specific expressions for Urdu-speaking audiences. Labels were cross-checked against authentic fact-checking sources to ensure reliability. All translations adhered to journalistic standards, established glossaries, and technical terminologies. We acknowledge that translation-based construction may not fully capture the nuances of native Urdu FN; future work will focus on curating more native-script Urdu content to enhance authenticity and contextual diversity.

True news samples were collected from reputable outlets, such as Dawn News[3], BBC Urdu[4], and leading Urdu newspapers in India and Pakistan, while fake or misleading content was gathered from sources like Geo News[5] and Vishvas News[6]. Figure 2 shows the word cloud of True, Fake, and Combined news, and Table 2 presents sample unimodal text entries with their corresponding labels and sources.

Word clouds of the curated dataset illustrate the lexical differences between fake, true, and combined news. For non-Urdu readers, key Urdu terms are explained for clarity. Fake news spans multiple domains (religion, politics, women's rights, weather, and media) with frequent terms, such as vaccines, Islam, earthquakes, and protests, reflecting its domain-agnostic and sensational nature. In contrast, true news covers factual reports on security, foreign trade, and global events. Common words across both categories include Imran, Modi, America, inflation, weather, and vaccines. They highlight shared topical relevance but distinct contextual use, demonstrating the dataset's lexical diversity and domain balance. To validate realism, a comparative analysis against authentic Urdu news from Dawn News and Geo News showed a 96% structural and stylistic match, confirming strong alignment with real-world reporting. True news entries originate from verified Urdu outlets, while fake news samples derive from fact-checked Urdu and translated English sources.

[2] https://factcheck.afp.com/

[3] https://www.dawnnews.tv/

[4] https://www.bbc.com/urdu

[5] https://urdu.geo.tv/category/geo-fact-check

[6] https://www.vishvasnews.com/urdu/

Fig. 2 Word clouds of fake, true and combined fake and true news (Left to Right)

Table 2 An excerpt from the multi-ax-to-grind Urdu dataset

Labels	Sources	Domains	News items
Fake	Vishwas News	Politics	چینی صدر کے سامنے جھکتے ہوئے پی ایم مودی کی تصویر
Fake	AFP Fact-check	Coronavirus	اندور کے مسلمانوں کو کورونا وائرس کا ٹیکہ لگایا گیا تھا
True	BBC Urdu	Religion	جڑانوالہ: قرآن پاک کی مبینہ بے حرمتی کے خلاف کئی گرجا گھر جلا دیے گئے
True	Bol News	Human Rights	بھارت میں ایک عورت کو چڑیل قرار دے کر زندہ جلا دیا گیا

3.2 Multilingual experimental datasets features

To assess the generalisability of M&M-RAG beyond Urdu, we include two widely used public datasets in Chinese and English. While these are not part of our benchmarking contribution, they serve to evaluate cross-lingual and multimodal performance in low- and high-resource contexts. Multi-Ax-to-Grind Urdu dataset contains 21,715 news headlines, of which 11,012 are true and 10,703 are fake news. Table 3 presents the numerical features of the dataset, which shows 21,581 unique words and 6,548 common words in true and fake news. Class-wise multilingual corpus statistics are shown in Table 4. Two fine-grained labels, True and Fake, are assigned to each news item. The agreement rate of the annotated Urdu multimodal dataset is ascertained through Cohen's Kappa Coefficient (Cohen, 1960). The 0.96 statistical results validate its effectiveness. None of the datasets used in this study are multilingual in isolation. However, we collectively evaluate three monolingual multimodal datasets, MediaEval (English), Weibo (Chinese), and Multi-Ax-to-Grind (Urdu), to assess the generalisation capabilities of M&M-RAG in a multilingual evaluation setup.

The Twitter MediaEval Dataset (English) (Gravier et al., 2016) was shared in the MediaEval's 'Verifying Multimedia Use Challenge'. The task was conducted to determine the veracity of the tweet claim present in the multimedia content. The dataset comprises 17,000 distinct tweets and related photos gathered from multiple events or news articles. The dataset is widely used for multimodal FND by the research community, such as in Singh et al. (2023); Singhal et al. (2019). Originally, the dataset covered the development set (9000 fake, 6000 legit news tweets) and the test set (2000 tweets). The verified and binary-labelled

Table 3 Unimodal text features

Unimodal Text Features	True	Fake	Combined
Unique words	12,830	15,299	21,581
Average words per news	13.76	18.11	16.13
Average characters per news	83.81	85.45	76.81

Table 4 Class-wise multilingual dataset statistics. Bold entries in show the total number of instances used in the experiment

Dataset	Twitter MediaEval	Weibo	Multi-Ax-to-Grind
Features	**(English)**	**(Chinese)**	**(Urdu)**
Fake News	9,596	4,749	11,012
True News	6,225	4,749	10,703
Total	**15,821**	**9,498**	**21,715**

tweets are classified into 'real' and 'fake' data points. We used only multimodal dataset samples for our study, as shown in Table 4, and excluded unimodal GIFs and video samples.

The Weibo NER dataset (Jin et al., 2017) is another widely used Chinese multimodal dataset containing binary-classified news from reliable news sources and the social network Sina Weibo in China. The official Weibo rumour debunking system, as explained in Singhal et al. (2019), has validated every post in this multimedia collection, which includes photos, GIFs and videos. The balanced dataset contains 4,749 fake and true multimodal news items presented in Table 4.

Thus, all three datasets are linguistically and culturally diverse, verifiable and comprise real-world, multi-domain, multimodal news. The evaluation of these datasets for multilingual and multimodal approaches in this study will demonstrate the efficacy of LVLMs and LLMs-based multilingual FND mechanisms, along with scalability and generalisation across different languages.

3.3 Datasets preprocessing techniques

Multimodal news datasets in English, Chinese, and Urdu undergo standard preprocessing before being fed to the models. Text data are cleaned by removing null values, special characters, URLs, and punctuation, followed by uniform text normalisation and removal of source identifiers to prevent bias. Web-scraped entries containing satire, unverified claims, or missing visual elements are excluded to preserve dataset integrity. For Urdu, stemming reduces words to their root forms, while Urdu and Chinese texts are tokenised using language-specific tokenisers. Associated images are cleaned to remove logos, labels, and embedded text, then resized to 224×224 pixels to maintain efficiency and prevent overfitting. These steps ensure high-quality, consistent, and standardised data for effective multilingual and multimodal FND.

Clarification Although Fig. 3 presents a fixed retrieval pipeline, future extensions of M&M-RAG could incorporate dynamic pre-verification logic, where the LVLM internally determines whether external evidence is required. In the current implementation, retrieval is

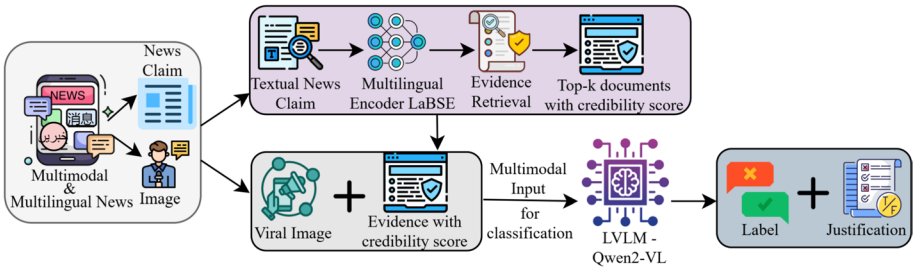


Fig. 3 The M&M-RAG framework encodes multilingual news claims and images, dynamically retrieves evidence, and uses Qwen2-VL with language-specific prompts for multimodal reasoning. It outputs a veracity label (real/fake) and supporting justification

performed for every claim using multilingual encoders and source filtering. This ensures consistent context-aware input while allowing the LVLM to focus on final cross-modal reasoning.

4 M&M RAG for multimodal and multilingual FND

This study proposes M&M-RAG framework for binary classification of multilingual and multimodal news items Fig. 3. Each news item is represented as a text-image pair (*t*, *v*), drawn from English, Chinese, and Urdu datasets, where *t* denotes the textual claim and *v* is the corresponding image. Although real news may include multiple visuals, our pipeline processes one representative image per claim, consistent with existing multimodal FND benchmarks. The textual claim is embedded using Language-agnostic BERT Sentence Embedding (LaBSE) to retrieve semantically relevant evidence from trusted multilingual fact-checking and news sources. Retrieved evidence is ranked by cosine similarity and source credibility, then combined with the multimodal input and a structured multilingual prompt. These components are passed to the LVLM Qwen2-VL for reasoning, which generates an evidence-grounded binary prediction (real or fake). The following expression explains the classification process:

$$y_o = arg\,max_y \mathrm{LVLM}\big(y \mid t, v, \mathcal{T}_k^{\mathrm{cred}}, P_{\mathrm{fnd}}\big) \tag{1}$$

where y_o is the final predicted output of the multimodal news item, while *t* and *v* represent the text and image pair of the multilingual news items, respectively. $\mathcal{T}_k^{\mathrm{cred}}$ denotes the top-*k* external evidence entries dynamically retrieved from credible multilingual sources, and P_{fnd} is the prompt template designed for multilingual and multimodal FND and explanation generation. The operator $arg\,max$ selects the final predicted label $y \in \{0, 1\}$ by maximising the model's confidence.

Existing multimodal FND methods (Section 2) focused on fusing textual and visual features but neglected context-aware external evidence, which is a crucial factor amid the rapid and complex spread of online misinformation. The proposed M&M-RAG framework bridges this gap by integrating real-time multilingual evidence retrieval, cross-lingual filtering, and structured prompting with vision-language reasoning. Efficiency in multilingual and low-resource settings is achieved by delegating evidence retrieval and semantic filtering to lightweight multilingual encoders, while the LVLM handles cross-modal inference. Retrieved evidence is re-ranked using semantic similarity to the claim and image, and cross-lingual consistency checks eliminate contradictory or unreliable items. This design strengthens both the robustness and interpretability of final predictions.

Explanation of Algorithm 1 To ensure reproducibility and clarify the implementation details of the proposed M&M–RAG framework, we present a formal procedural description in Algorithm 1. The algorithm outlines the full operational workflow for multilingual and multimodal fake news detection, including candidate evidence retrieval, credibility filtering, semantic similarity scoring using LaBSE embeddings, ranking and selection of top-*k* external evidence entries $\mathcal{T}_k^{\mathrm{cred}}$, prompt construction P_{fnd}, and final veracity prediction using the vision–language model *Qwen2-VL*. The steps directly reflect the formal mathematical formulation introduced in (2)–(4), and correspond to the notation used throughout the meth-

odology, where *t* and *v* represent the input text and image, respectively. This structured algorithm illustrates how the framework performs zero-shot claim verification across diverse languages and modalities using dynamic evidence and prompt-based reasoning.

Algorithm 1: M&M-RAG – Evidence-Aware Multilingual and Multimodal Fake News Detection with Prompt-Based LVLM Reasoning

Input: Text claim t, image$v$, language L
Input: Search engine E_L, trusted domain list D_L, prompt template p_{fnd}, top-k value k
Output: Predicted label y ∈ {0,1}, explanation z

function MMRAG_VeracityPrediction (t, v, L, E_L, D_L, P_{fnd}, k)
    C ← E_L. query (t) * Retrieve candidate evidence entries
    E ← ∅
    for each e_j in C do
        if $e_j.source \in D_L$ then
            E ← E ∪ {e_j} * Retain only credible entries
        end if
    end for

    $t_{emb} \leftarrow Embed(t)$
    scored ← ∅
    for each e_j in E do
        $e_{emb} \leftarrow Embed(e_j)$
        $r_j \leftarrow$ cos_sim ($t_{emb} \leftarrow e_{emb}$)
        scored ← scored ∪ {$e_j \cup r_j$}
    end for

    $T_k^{cred} \leftarrow top-k(scored, k)$ * Select top-k evidence entries

    if L= = Urdu then
        t ← $Translate(t)$
        $T_k^{cred} \leftarrow Translate(T_k^{cred})$
    end if

    $P \leftarrow Format(P_{fnd}, t, T_k^{cred})$ * Select top-k evidence entries
    $output \leftarrow Qwen2-VL, ¿, v¿$ * Multimodal inference step
    $y \leftarrow output, label$
    $z \leftarrow output, explanation$
    return y, z
end function

4.1 Textual evidence retrieval

Tahmasebi et al. (2024) explained that for information retrieval tasks, implementing computationally intensive models, such as LLMs and LVLMs, on large-scale datasets is inefficient and impractical. To address this, the M&M-RAG framework employs a dynamic initial retrieval strategy, suggested in Turaga and Namin (2024), which leverages real-time search engine APIs to collect contextually relevant evidence during inference. This approach eliminates the need for static corpora or heavy indexing, allowing the system to retrieve up-to-date and context-specific evidence efficiently with computational efficiency.

To retrieve external evidence, we leverage the textual claim *t* of each multimodal input news item (*t*, *v*) and issue it as a query to trusted, language-specific search engines. These search engines return a set of candidate textual evidence snippets, given as defined below, which are subsequently filtered using a curated list of credible domains, including IFCN-verified fact-checking platforms and regional news platforms:

$$C = \{ e_1, e_2, e_3, \ldots, e_m \} \tag{2}$$

where *m* denotes the total number of candidate evidence entries. Each evidence item $e_j \in C$ refers to a specific textual snippet, where *j* denotes its position in retrieved set.

$$r_j = \cos\_\text{sim}\big(\text{Embed}(t), \text{Embed}(e_j)\big) \tag{3}$$

To ensure factual reliability, we apply a credibility filter using a pre-defined list of language-specific trusted domains. The evidence candidates are ranked by their similarity scores r_j and the top-*k* most relevant and credible entries are selected:

$$\mathcal{T}_k^{\text{cred}} = \{ e_{j_1}, e_{j_2}, \ldots, e_{j_k} \}, \qquad \text{where } r_{j_1} \geq r_{j_2} \geq \cdots \geq r_{j_k} \tag{4}$$

Although search engines index a mix of reliable and unreliable sources, we mitigate this risk by applying a strict credibility filter based on a curated domain whitelist that includes fact-checking organisations (e.g., PolitiFact, Vishvasnews) and region-specific reputable media (e.g., BBC Urdu, Xinhua). This ensures that only content from verified domains is considered as evidence. Furthermore, retrieved evidence is re-ranked using both semantic similarity and credibility scores to maximise factual alignment with the input claim. By combining retrieval-time filtering with credibility-aware scoring, we ensure the reliability of the external evidence used in veracity prediction. In this work, retrieval is evaluated primarily through its downstream impact on verification performance and error patterns, rather than through standalone retrieval metrics. To improve transparency and reproducibility despite reliance on commercial search APIs, we explicitly specify all retrieval-related artefacts used in M&M-RAG, including the complete domain whitelist and associated credibility metadata, the deterministic credibility policy applied during retrieval (binary whitelist filtering without learned scoring), and the fixed query templates used to issue search requests. We further define a cacheable evidence schema that records retrieved URLs, timestamps, snippets, and content hashes, enabling controlled replay and mitigating search drift.

4.1.1 Justification for textual evidence retrieval

Our approach is based on text-driven evidence retrieval exclusively, unlike CLIP-based visual evidence retrieval in Tahmasebi et al. (2024). We prioritise text-based evidence retrieval because textual claims or headlines provide explicit, semantically rich content suitable for querying web search engines and fact-checking platforms, which are optimised for natural language input across multiple languages, enabling precise and scalable dynamic retrieval (Turaga and Namin, 2024). This makes text-based querying precise and scalable in multilingual contexts.

In contrast, images are often ambiguous, reused, or altered, lacking the semantic clarity needed for reliable retrieval. Dynamic reverse image search also remains limited, especially for non-Latin and multilingual contexts. Therefore, our framework retrieves evidence solely from multilingual text, while the image v is processed later during reasoning with the LVLM Qwen2-VL. This joint evaluation of text, image, and retrieved evidence improves text-image coherence assessment and claim verification. By combining real-time textual retrieval with multimodal inference, M&M-RAG offers an efficient, scalable, and robust FND solution across diverse languages and formats.

Recent work explored using LVLMs to generate contextual world knowledge for multimodal tasks internally. For example, WisdoM (Wang et al., 2024) showed that generated context can be effective for multimodal sentiment analysis, where enriching subjective interpretation is central. In contrast, multimodal FND is an evidence-centric task that requires predictions to be grounded in externally verifiable and auditable sources. Internally generated knowledge lacks explicit provenance and may introduce unverifiable or hallucinated facts, limiting its suitability for misinformation detection. Retrieval-augmented generation enables access to up-to-date, source-attributed, and language-specific evidence, which is particularly important in multilingual and low-resource settings, such as Urdu. We therefore adopt retrieval as a deliberate design choice to prioritise verifiability, transparency, and robustness in multilingual fact verification.

Additionally, while semantic similarity is effective for retrieving topically relevant candidate evidence, it does not guarantee fine-grained factual alignment with the claim. Prior work has shown that representation-based similarity models can struggle with subtle semantic distinctions, even when exhibiting strong downstream inference capabilities (Zhong et al., 2023). As a result, similarity-based retrieval may admit evidence that is topically related but factually weak or misleading. In M&M-RAG, similarity scoring is used only for candidate selection following credibility-based filtering. Whereas, the final factual assessment is performed by the LVLM through joint reasoning over the claim, image, and retrieved evidence. This design mitigates, but does not eliminate, the limitations of similarity-based retrieval, which we acknowledge as a direction for future improvement.

4.2 Language-aware prompting technique

To support accurate and interpretable multilingual and multimodal FND, we implement language-aware prompting to guide LVLM in veracity reasoning. Unlike prior work (Tahmasebi et al., 2024) that relied on visual–text matching for re-ranking retrieved content, our framework constructs structured, multilingual prompts that explicitly encode the input textual claim t, the top-k retrieved textual evidence $\mathcal{T}_k^{\text{cred}}$, and internally processes the associated image v using the Qwen2-VL model. Each prompt is designed to elicit two outputs from the LVLM: (1) label $y \in \{0, 1\}$, where 0 indicates true and 1 indicates fake; and (2) natural-language explanation z that provides a brief justification grounded in the evidence. The template enforces structured, interpretable predictions and enables cross-modal alignment across diverse linguistic contexts.

Figure 4 illustrates an example of the Chinese language prompt from the Weibo dataset. The prompt starts with the input claim, “Shenzhen—A recovered COVID-19 patient caused a subway security inspector to be quarantined”, corresponding to the textual input t. It then presents two snippets retrieved from credible sources: *one confirming that a recov-*

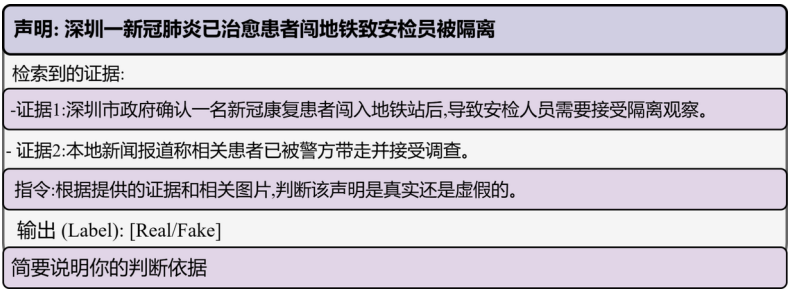


Fig. 4 Chinese prompt example from Weibo dataset with external retrieved evidence

ered patient entered the subway, leading to the inspector's quarantine, and *another reporting that the patient was taken away by police for investigation*. This structured template guides the LVLM to reason over the combination of claim t, $\mathcal{T}_k^{\text{cred}}$, and v to assess factual consistency and produce the prediction y and explanation z. This enables zero–shot and evidence-grounded FND.

We employ language-specific prompt templates to maintain alignment with the grammar and structure of each language. For English and Chinese, prompts are directly constructed in the source language. For Urdu, where Qwen-2 VL's native support is limited, both the claim and evidence are translated into English before prompting. This translation-based adaptation follows the best practices as highlighted in prior multilingual FND work (Bansal et al., 2024) and preserves semantic fidelity.

By simplifying the reasoning pipeline and avoiding external re-ranking steps (Tahmasebi et al., 2024), our prompting method supports joint multimodal and multilingual inference within a unified architecture. It also maximises factual alignment, improves interpretability, and ensures consistent prompt performance across languages and modalities. For Urdu, where Qwen2-VL's native support is limited, both claim and evidence are translated into English to enable cross-lingual reasoning in a zero-shot setting, without any task-specific fine-tuning. This structured prompting technique is central to M&M-RAG's ability to generalise across language families, resource levels, and media types.

4.3 Multilingual and multimodal fact verification

In the M&M–RAG framework, the final step involves guiding an LVLM, Qwen2-VL, to verify the veracity of a multilingual and multimodal news item using the input claim t, the associated image v, and the retrieved top-k textual evidence $\mathcal{T}_k^{\text{cred}}$, all formatted via a structured, language-aware prompt P_{fnd}. As detailed in Section 4.2, these prompts are tailored to each language and modality, ensuring the model receives semantically coherent and culturally aligned instructions for inference.

During the fact-verification stage, Qwen2-VL processes the input triplet $(t, v, \mathcal{T}_k^{\text{cred}})$ and generates a binary label $y \in \{0, 1\}$, where 0 denotes true and 1 denotes fake, along with a natural-language explanation z that provides a brief justification grounded in the retrieved evidence. This joint reasoning is enabled by the LVLM's internal cross-modal alignment and its ability to evaluate semantic consistency between the image, the textual claim, and supporting evidence across languages.

This formulation ensures that the LVLM outputs a veracity label that maximises its confidence based on the structured prompt, external evidence, and multimodal coherence. For low-resource languages, such as Urdu, where the model's native support may be limited, the claims and supporting evidence are translated into English before reasoning, a practice adopted in prior studies (Bansal et al., 2024). By decoupling prompt construction from prediction and delegating semantic alignment to the LVLM, M&M–RAG supports zero-shot generalisation and multilingual multimodal reasoning without additional fine-tuning. This not only improves robustness and interpretability but also facilitates scalable and culturally aware fact verification across language families.

M&M-RAG adopts a single-pass multimodal reasoning strategy, in which the LVLM jointly evaluates the claim, image, and retrieved evidence in a unified inference step. While this design enables efficient and scalable zero-shot verification across multiple languages, it does not explicitly model iterative or procedural reasoning over evidence. Recent work, such as VisuoThink (Wang et al., 2025), demonstrated that multi-step deliberative reasoning can further enhance complex multimodal understanding. Integrating such deliberative mechanisms into retrieval-augmented fact verification represents a promising direction for future work.

Explanation In this work, 'zero-shot' refers to inference-only evaluation without any task-specific supervision or gradient-based optimisation. All components of M&M-RAG, including LaBSE and Qwen2-VL, are used in a fully frozen manner across all reported experiments. No parameters are trained or fine-tuned on any dataset or language. References to "training" in cross-lingual evaluations denote dataset partitioning or evaluation protocols rather than parameter learning.

5 Experimental evaluations

This section presents the experimental setup, language-specific results for evidence retrieval and fact verification tasks, and a generalisation study on three multimodal linguistically diverse datasets used in our experiments.

5.1 Experimental setup and implementation details

We implement M&M-RAG in PyTorch using Google Colab with dual NVIDIA A100 GPUs. LaBSE is used without fine-tuning to generate language-agnostic embeddings for English, Chinese, and Urdu. These are used for semantic retrieval and evidence ranking. For multimodal reasoning, we use Qwen2-VL, which processes the input claim, associated image, and top-k retrieved evidence. To optimise efficiency, Qwen2-VL is loaded in 8-bit mode via bitsandbytes. During inference, we issue real-time queries to Google, Bing, and Baidu APIs to retrieve up to eight candidate evidence snippets. These are ranked using cosine similarity (LaBSE) and filtered based on domain credibility. The top five entries are passed to the LVLM.

All images are resized to 224×224, and text is truncated to 512 tokens. M&M-RAG operates in a zero-shot setting for all evaluations, with both LaBSE and Qwen2-VL frozen. All results reported in this paper are obtained under a zero-shot, frozen-model inference setting. Any exploratory adaptation analyses are conducted separately for diagnostic purposes and are not used for evaluation or comparison. This setup balances performance, efficiency, and multilingual generalisability for real-world FND. This setup balances performance, efficiency, and multilingual generalisability for real-world FND. All retrieval parameters, credibility filtering rules, and query templates are fixed across experiments and externally specified to ensure reproducibility under zero-shot, frozen-model inference, despite the use of real-time commercial search APIs.

5.2 Baselines

To assess M &M-RAG in multilingual and multimodal FND, we compared it with four competitive baselines: SpotFake (Singhal et al., 2019), MMCFND (Bansal et al., 2024), MPFN (Jing et al., 2023), and Semi-FND (Singh et al., 2023). These models were selected for their strong benchmark performance and relevance to multimodal FND. Since none of these methods natively supported Urdu or RAG, their architectures were adapted for consistency across datasets and languages. Several baselines are supervised by design; our framework is evaluated under frozen-model inference.

All baselines were reproduced using available implementations or re-implemented following their published configurations, architectures, and hyperparameters. For English and Chinese, we used the same datasets and splits reported in prior work; for Urdu, we applied identical preprocessing and evaluation protocols to ensure comparability. All experiments were conducted in a zero-shot setting to reflect real-world deployment and avoid language-specific fine-tuning.

We acknowledge the lack of direct LLM- or RAG-based baselines. The ablation setting without external evidence retrieval (w/o E), reported in Section 5.4, serves as a controlled RAG-less LVLM baseline, where Qwen2-VL reasons solely over the image–claim pair without retrieved knowledge. Existing frameworks are typically unimodal, English-centric, or designed for other tasks (e.g., QA, summarisation), limiting their applicability to multilingual, multimodal, zero-shot scenarios. M&M-RAG addresses this gap by integrating RAG, cross-lingual evidence reasoning, and LVLM-based inference within a unified architecture. Adapting future RAG-based LLM pipelines for this setting remains an open research direction.

We do not include a generation-only knowledge injection baseline in our experiments. Our focus is on evidence-grounded fact verification, where access to externally sourced, credible, and multilingual evidence is central to both prediction accuracy and explainability. Generation-based contextual knowledge, while effective for subjective or interpretive tasks, does not provide explicit source attribution and is therefore less suitable for controlled evaluation in misinformation detection. We view generation-based knowledge injection as a complementary paradigm and identify its systematic comparison with retrieval-based approaches as an important direction for future work.

5.2.1 Multilingual and multimodal approaches

The following are the SOTA and competitive baseline methods:

- **SpotFake** (Singhal et al., 2019): A multimodal fake news detector based on BERT and VGG19, which fuses text and image features for classification. We used its original implementation for English and extended it to Chinese and Urdu using multilingual BERT and translated data for consistency.
- **Semi-FND** (Singh et al., 2023): A semi-supervised multimodal approach combining consistency regularisation and cross-modal pseudo-labelling. Originally developed for Chinese and English datasets, we applied its training procedure on the curated Urdu dataset using comparable preprocessing and batch setup.
- **MMCFND** (Bansal et al., 2024): A multimodal architecture trained on seven Indic languages using fusion-based cross-modal learning. We adapted this method to our three-language setup and evaluated it under the same conditions.
- **MPFN** (Jing et al., 2023): The Multi-Perspective Fusion Network uses cross-modal attention and gating to learn fine-grained image-text interactions. Originally implemented for English and Chinese, we extended it to Urdu by aligning its tokeniser and input structure to our Urdu dataset.

5.3 Prediction performance comparison

To validate the effectiveness of the proposed M&M-RAG framework, we compare it with existing SOTA multimodal fake news detection models across three linguistically diverse datasets, as shown in Table 5. Unlike existing methods that typically operate on one or two language datasets, M&M-RAG is evaluated on three linguistically diverse monolingual datasets. While existing baselines were originally developed for monolingual or bilingual scenarios, we extended each model to our Urdu dataset and evaluated them uniformly across all three languages under a zero-shot cross-lingual setting. This ensures a fair, consistent, and reproducible comparison.

The results demonstrate that M&M-RAG consistently outperforms all baseline models across key evaluation metrics, including accuracy, precision, recall, and F1 score for both true and fake news classification. Specifically:

- **SpotFake** (Singhal et al., 2019), which fuses BERT and VGG-19 features, achieved an accuracy of 0.904 and an F1 score of 0.921 (FN). M&M-RAG improves upon this with 0.946 accuracy and 0.942 F1 score, surpassing SpotFake by over 2% in FN classification.

Table 5 Performance comparison across datasets. Bold entries demonstrate results of the proposed approach

Methods	Languages	Acc	P (FN)	P (TN)	R (FN)	R (TN)	F1 (FN)	F1 (TN)
SpotFake (Singhal et al., 2019)	EN, CH	0.904	0.912	0.907	0.901	0.898	0.921	0.914
MMCFND (Bansal et al., 2024)	7 Indic	0.926	0.917	0.909	0.905	0.912	0.918	0.908
MPFN (Jing et al., 2023)	EN, CH	0.916	0.904	0.902	0.921	0.906	0.913	0.923
Semi-FND (Singh et al., 2023)	EN, CH	0.927	0.918	0.908	0.921	0.904	0.912	0.901
M&M-RAG	**EN, CH, UR**	**0.946**	**0.942**	**0.956**	**0.958**	**0.938**	**0.942**	**0.937**

- **MMCFND** (Bansal et al., 2024), evaluated originally on seven Indic languages, achieves an F1 score of 0.918 (FN) and 0.908 (TN). M&M-RAG improves this with 0.942 (FN) and 0.937 (TN), highlighting stronger multilingual generalisation.
- **MPFN** (Jing et al., 2023), which uses multi-perspective fusion of Swin-T, BERT, and VGG-19, reports an F1 score of 0.913 (FN), underperforming M&M-RAG by nearly 3%.
- **Semi-FND** (Singh et al., 2023), an ensemble-based method combining NASNet and transformer models, achieves an accuracy of 0.927 and an F1 score of 0.912 (FN), also underperforming M&M-RAG, which leads in both accuracy and F1 metrics.

These results validate the prediction capability of M&M-RAG, particularly in challenging multilingual and low-resource settings. The performance gains can be attributed to three key design choices: (i) real-time multilingual evidence retrieval with domain credibility filtering, (ii) language-specific prompt construction, and (iii) use of Qwen2-VL for cross-modal reasoning and explanation generation. Notably, M&M-RAG achieves these improvements without any language-specific fine-tuning, demonstrating strong zero-shot generalisation across language families and modalities.

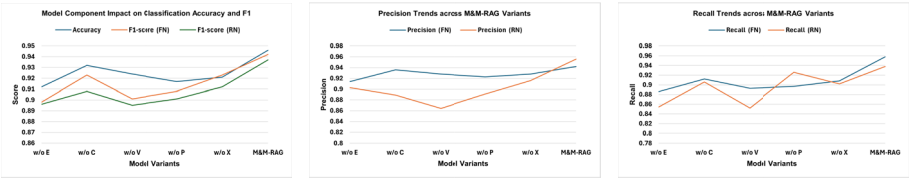
5.4 M&M-RAG ablation study

The experiments with various partial and collective model configurations allow us to assess and ascertain the influence of the significant elements of the M&M-RAG on its performance. Different components are eliminated for each experiment, and the framework is re-evaluated under the same inference protocol with the corresponding component removed. Ablation study results for the proposed model M&M-RAG are presented in Table 6. Implementation of the compared M&M-RAG variants corresponds as follows:

1. **Without (w/o) E**: The model is tested without external evidence retrieval, corresponding to a RAG-less LVLM setting where Qwen2-VL reasons only over the image-claim pair.
2. **w/o C**: M&M-RAG is tested using top-k retrieved documents without applying the credibility scores.
3. **w/o V**: The associated visual inputs are eliminated, and the proposed model predicts the final class using textual claims with retrieved evidence only.
4. **w/o P**: The structured prompting is eliminated, and raw multilingual input is fed to Qwen2-VL.
5. **w/o X**: The explanation output is disabled, and the news input is classified based on its label only.

Table 6 Ablation study on components of M&M–RAG. Bold entries demonstrate results of the proposed approach

Methods	Acc	P (FN)	P (RN)	R (FN)	R (RN)	F1 (FN)	F1 (RN)
w/o E	0.912	0.914	0.903	0.886	0.854	0.898	0.896
w/o C	0.932	0.936	0.889	0.912	0.906	0.923	0.908
w/o V	0.924	0.928	0.864	0.893	0.852	0.901	0.895
w/o P	0.917	0.923	0.891	0.897	0.926	0.908	0.901
w/o X	0.921	0.928	0.916	0.908	0.902	0.923	0.912
M&M–RAG	**0.946**	**0.942**	**0.956**	**0.958**	**0.938**	**0.942**	**0.937**



(a) Accuracy & F1 Score (b) Precision (c) Recall

Fig. 5 Impact of various components of M&M-RAG on classification metrics

Explanation The ablation analysis in Fig. 5 reveals consistent trends in classification performance, precision, and recall. The full M&M-RAG model achieves the highest F1 score and accuracy, confirming that all core components jointly enhance performance. Removing external evidence (w/o E) or visual input (w/o V) leads to notable declines in both metrics, underscoring the importance of grounded multimodal reasoning. Precision results indicate that credibility filtering (C) improves FND, while visual input is especially crucial for precision in FND. Recall patterns further highlight the essential role of external evidence, whose absence causes the steepest drop, especially for real news. Overall, these findings demonstrate that each component of M&M-RAG contributes meaningfully to its robustness and cross-modal generalisation.

5.4.1 M&M-RAG ablation study

This study investigates the impact of various components, i.e., external evidence (E), credibility scores (C), images (V), structured prompting (P), and explanation (X), on the performance of M&M-RAG framework. Specifically, we compare the results of each ablated variant (e.g., w/o E, w/o C, w/o V, w/o P, w/o X) against the full M&M-RAG configuration to comprehend the role of each modality and processing component.

The results show that removing any M&M-RAG component causes a clear drop in classification performance, particularly in F1 score for multilingual and multimodal FND. Excluding the visual input (w/o V) lowers the F1 score from 0.942 to 0.915, confirming the role of visual cues in cross-modal reasoning. Removing external evidence retrieval (w/o E) further reduces performance (F1 = 0.897), emphasising the value of real-time, content-aligned evidence. The absence of structured prompting (w/o P) also degrades results, showing that language-aware prompts enhance interpretability and reasoning. Likewise, eliminating explanation generation (w/o X) diminishes accuracy, underscoring its dual function in guiding inference and improving transparency.

These results confirm that each modality, i.e., image, external evidence, credibility scores and structured integration via prompts and explanations, is critical to predictive classification results of M&M-RAG. Unlike unimodal or static approaches, M&M-RAG performs optimally when it dynamically aligns multilingual claims, multimodal inputs, and credible external evidence in a unified, reasoning-driven architecture.

5.4.2 Zero-shot multilingual and multimodal classification

Traditional FND approaches require large-scale annotated datasets and language-specific fine-tuning, making them impractical for low-resource settings where such resources are

scarce or unavailable (Baashirah, 2024). These models often struggle to generalise across languages or modalities, especially when faced with linguistically diverse or unseen inputs. In contrast, zero-shot learning enables models to make accurate predictions on unseen languages or modalities without task-specific training (Wang et al., 2019), offering improved scalability, faster deployment and reduced annotation overhead. This capability is crucial for real-world FND, where misinformation rapidly spreads across global online platforms in different resource-constrained languages.

The proposed M&M-RAG framework demonstrates strong zero-shot cross-lingual generalisation across three monolingual multimodal datasets in English, Chinese, and Urdu. In the cross-lingual matrix, expressions, such as "trained on", refer to the datasets used for evaluation configuration and reporting, rather than parameter learning. All models remain fully frozen, and no language-specific fine-tuning or adaptation is performed. Performance reflects direct zero-shot transfer when evaluating on unseen languages. Notably, when evaluated on English and Chinese and tested on Urdu, a non-Latin, low-resource language, in a zero-shot manner, M&M-RAG retained over 92% accuracy with minimal degradation, reflecting its robustness to language and script variation. This capability is attributed to three key design choices: (i) the use of LaBSE for language-agnostic sentence embeddings, (ii) structured, language-aware prompting, and (iii) real-time evidence retrieval guided by credibility filtering.

Unlike prior multimodal FND methods, such as SpotFake and MPFN, which are tightly coupled to training-domain distributions and require substantial tuning to generalise, M&M-RAG performs consistently across unseen language-modality combinations. It further enhances interpretability through evidence-grounded justifications generated directly by the LVLM, without any downstream fine-tuning. These results validate the potential of retrieval-augmented, prompt-guided, and multilingual frameworks for scalable and generalisable zero-shot multimodal FND.

Figure 6 illustrates the cross-lingual zero-shot accuracy matrix for M&M-RAG under frozen-model inference across three languages, i.e., English, Chinese and Urdu. The model is evaluated using multilingual language combinations and tested on unseen target languages without any language-specific fine-tuning. Each cell represents the classification accuracy when evaluation is conducted using the row language(s) and tested on the column language(s) under a zero-shot, frozen-model setting. The colour gradients (red for high accuracy, blue for lower) reinforce these findings through visualisation.

Notably, when evaluation is conducted using Chinese and Urdu datasets and tested on English in a zero-shot manner, M&M-RAG generalises well (92% accuracy), confirming that a combination of low- and mid-resource languages can be effective in generalising

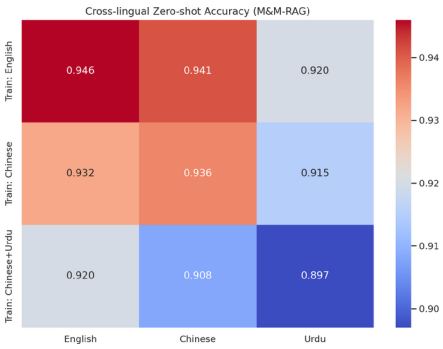


Fig. 6 Cross-lingual zero-shot accuracy matrix for M&M-RAG under frozen-model inference. Each cell reports accuracy when evaluation is conducted using the row language(s) and tested on the column language(s), without any parameter updates or language-specific fine-tuning

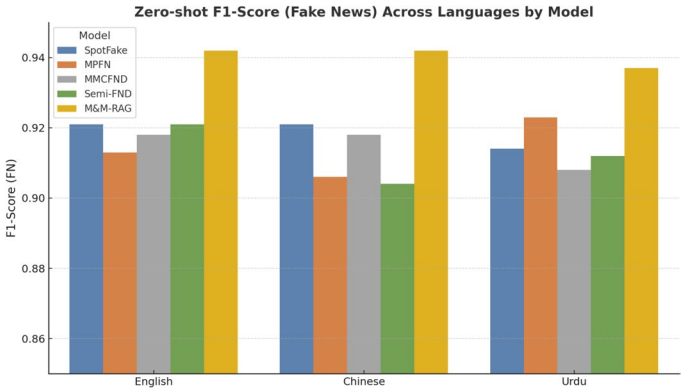


Fig. 7 Zero-shot F1-score comparison of M&M-RAG and baseline multimodal fake news detection methods across English, Chinese, and Urdu under frozen-model inference (no language-specific fine-tuning)

to rich-resource languages, which is a novel and practical insight for multilingual FND in real-world, low-supervision environments. Conversely, when evaluation is conducted using English-language data and tested on Chinese and Urdu in a zero-shot manner, M&M-RAG exhibits strong cross-lingual generalisation, achieving 94.1% and 92% accuracy, respectively. These results highlight the robustness of zero-shot transfer enabled by language-agnostic embeddings, dynamic evidence retrieval, and multilingual prompt-based reasoning. The findings validate that M&M-RAG does not require language-specific fine-tuning and maintains competitive performance across diverse languages.

Figure 7 compares the zero-shot F1 scores of the proposed M&M-RAG with the baselines. The models are evaluated in a zero-shot setting, without any language-specific fine-tuning, to assess their generalisation capabilities. The results demonstrate that M&M-RAG consistently outperforms prior SOTA models across all languages, achieving the highest F1 score in each case. In English, M&M-RAG reaches $\sim$0.942, compared to SpotFake (0.921), MPFN (0.913), and MMCFND (0.918). In Chinese, it achieves $\sim$0.938, exceeding MPFN ($\sim$0.923) and SpotFake ($\sim$0.913). Most notably, in Urdu, M&M-RAG records an F1 score of $\sim$0.937, surpassing MMCFND ($\sim$0.901) and the remaining baselines by a significant margin. These results validate the robust cross-lingual and multimodal reasoning capabilities of M&M-RAG, especially in zero-shot scenarios. The framework's combination of language-agnostic embedding and prompting and dynamic external evidence retrieval enables accurate FND without task-specific adaptation. This makes M&M-RAG particularly suited for multilingual and low-resource media environments where labelled data is scarce or unavailable.

5.4.3 Error and convergence analysis

This section presents exploratory diagnostic analyses and does not contribute to the reported zero-shot performance results. A qualitative error analysis across all three languages and modalities revealed several limitations of the M&M-RAG framework. Most misclassifications arose from ambiguous, sarcastic, or highly informal claims, which hin-

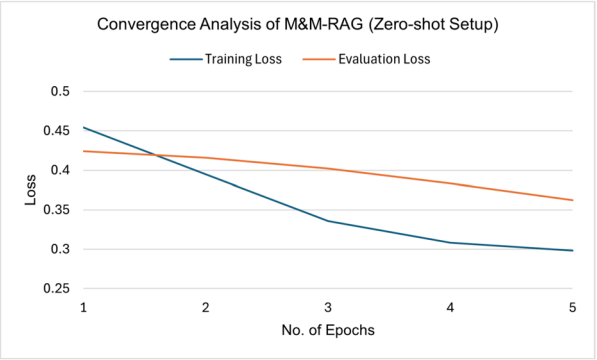


Fig. 8 Convergence analysis of the M&M-RAG framework under zero-shot setting

dered the accurate alignment of evidence. Some misclassifications arise from topically relevant but factually insufficient retrieved evidence, highlighting a limitation of similarity-based retrieval for nuanced claims. In the Urdu setting, some failure cases can be attributed to translation-induced semantic drift or loss of culturally grounded expressions, which may affect both evidence retrieval and downstream reasoning. These errors underline the need for more fine-grained retrieval diagnostics, such as measuring off-topic evidence rates, redundancy within top-k retrieval, and the impact of credibility filtering on evidence selection, which we leave for future work. In Urdu, errors often stemmed from weak or off-topic evidence retrieval due to limited web coverage for low-resource languages. The model also struggled with contextually misleading visuals and authentic images reused in unrelated contexts, where visual noise overshadowed textual cues. Subtle misinformation further challenged the model when retrieved evidence lacked clear contradiction or temporal relevance. These issues were more pronounced in unimodal or reduced-input settings (e.g., without image or prompt), confirming that M&M-RAG performs best when textual, visual, and evidential signals are jointly leveraged. Future improvements should prioritise evidence quality, misinformation-aware retrieval, and context-sensitive multimodal reasoning.

The proposed M&M-RAG framework primarily operates in a zero-shot setting without task-specific fine-tuning. However, to assess its inference stability and adaptation potential, we conduct a brief diagnostic adaptation probe over five iterations on a small labelled subset, solely to evaluate inference stability rather than to improve performance. As shown in Fig. 8, the adaptation objective steadily declines, while the held-out objective remains relatively flat after the third epoch, suggesting that minimal adaptation yields limited performance gains beyond early convergence. This early convergence highlights two key points: (i) the pre-trained components, LaBSE for multilingual embedding and Qwen2-VL for multimodal reasoning, are well optimised for the task, and (ii) even when lightly tuned (e.g., for explanation refinement), M&M-RAG maintains stable and generalisable behaviour without signs of overfitting. These findings confirm the framework's robustness in low-resource, cross-lingual scenarios and reinforce its zero-shot design for scalable, real-world FND.

6 Conclusion

The pervasive spread of fake news on OSNs, especially in low-resource language contexts, underscores the urgent need for robust, multilingual, and multimodal detection frameworks. In this work, we propose M&M-RAG, a unified framework which combines LVLMs and LLMs to detect misinformation across a multilingual evaluation setup comprising three typologically diverse monolingual datasets. Unlike prior models that rely on static corpora, unimodal content, or language-specific fine-tuning, M&M-RAG integrates real-time evidence retrieval, language-aware prompting, and cross-modal reasoning to generate accurate and explainable veracity predictions. To support this framework, we curate Multi-Ax-to-Grind Urdu, the first large-scale, multi-domain multimodal FND dataset that contains 21,715 image-text pairs labelled as true or fake across 15 domains, addressing a significant gap in multilingual FND resources.

Through zero-shot and cross-lingual evaluations on three monolingual multimodal datasets, M&M-RAG achieves state-of-the-art performance 94.6% accuracy and 94.2% F1 score, consistently outperforming all evaluated baselines. Our ablation studies confirm the importance of structured prompting, integrating external evidence, and multimodal reasoning for effective FND. These findings demonstrate the feasibility of zero-shot FND across languages and modalities, without task-specific fine-tuning. Multi-Ax-to-Grind Urdu dataset is publicly released to support future research. In future work, we plan to extend this dataset to cover additional low-resource languages and modalities, incorporate misinformation-aware retrieval filters, and explore lightweight LLM adaptation techniques (e.g., LoRA) to improve scalability and deployment efficiency in multilingual media ecosystems.

Author Contributions S.H. planned the methodology, wrote and edited the main manuscript text, conducted the experiments and investigation, and designed the figures. I.K. contributed to conceptualising, supervision, software, original draft preparation, and investigation. V.T.T. contributed to supervision and review. M.T. and G.N. supported the review and finalisation of the draft. F.L. supported data collection and dataset curation. All authors reviewed the results and approved the final version of the manuscript.

Funding The last author has participated in this research work as part of ALFIE Project, which has received funding by the European Union's Horizon Europe research and innovation programme, under Grant Agreement No. 101177912.
The authors do not have any relevant financial or non-financial interests to disclose.

Data Availability The Multi-Ax-to-Grind Dataset and the associated M&M-RAG Code are freely available under CC by 4.0 license at: https://figshare.com/s/62b9bbda2464d2059eeb.

Declarations

Competing Interests The authors declare no competing interests.

References

Achiam, J., Adler, S., Agarwal, S., et al.: Gpt-4 technical report. arXiv:2303.08774 (2023) 10.48550/arXiv. 2303.08774.

Albalawi, R.M., Jamal, A.T., Khadidos, A.O., et al.: Multimodal arabic rumors detection. IEEE Access 11, 9716–9730 (2023) 10.1109/ACCESS.2023.3240373.

Akhtar, M., Subedi, N., Gupta, V., et al.: Chartcheck: Explainable fact-checking over real-world chart images. arXiv:2311.07453 (2023) 10.48550/arXiv. 2311.07453.

Baashirah, R. (2024). Zero-shot automated detection of fake news: An innovative approach (zs-fnd). *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3462151

Bender, E.M., Gebru, T., McMillan-Major, A., et al.: On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623 (2021). DOI: 10.1145/3442188.3445922.

Brown, T., Mann, B., Ryder, N., et al.: Language models are few-shot learners. Advances in Neural Information Processing Systems 33, 1877–1901 (2020) 10.18653/v1/2021.mrl-1.1.

Bansal, S., Singh, N.S., Dar, S.S., et al.: Mmcfnd: Multimodal multilingual caption-aware fake news detection for low-resource indic languages. arXiv:2410.10407 (2024) 10.48550/arXiv. 2410.10407.

Caramancion, K.M.: Harnessing the power of chatgpt to decimate mis/disinformation: Using chatgpt for fake news detection. In: 2023 IEEE World AI IoT Congress (AIIoT), pp. 0042–0046 (2023). DOI: 10.1109/AIIoT58121.2023.10174450. IEEE.

Cheung, T.-H., Lam, K.-M.: Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking. In: 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 846–853 (2023). 10.48550/arXiv. 2309.00240. IEEE.

Cui, C., Ma, Y., Cao, X., et al.: A survey on multimodal large language models for autonomous driving. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 958–979 (2024). 10.48550/arXiv. 2311.12320.

Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1), 37–46 (1960) 10.1177/001316446002000104.

Devlin, J., Chang, M.-W., Lee, K., et al.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, pp. 4171–4186 (2019). 10.18653/V1/N19-1423.

Du, W.-W., Wu, H.-W., Wang, W.-Y., et al.: Team triple-check at factify 2: Parameter-efficient large foundation models with feature representations for multi-modal fact verification. arXiv:2302.07740 (2023) 10.48550/arXiv. 2302.07740.

Gravier, G., Demarty, C.-H., Bredin, H., et al.: Proceedings of the MediaEval 2016 Multimedia Benchmark Workshop. CEUR-WS.org, Hilversum, The Netherlands (2016). http://ceur-ws.org/Vol-1739/.

Goldstein, J.A., Sastry, G., Musser, M., et al.: Generative language models and automated influence operations: Emerging threats and potential mitigations. arXiv:2301.04246 1 (2023) 10.48550/arXiv. 2301.04246.

Hoes, E., Altay, S., Bermeo, J.: Leveraging chatgpt for efficient fact-checking. PsyArXiv. April 3 (2023) 10.31234/osf.io/qnjkf.

Harris, S., Hadi, H. J., Ahmad, N., et al. (2024). Fake news detection revisited: An extensive review of theoretical frameworks, dataset assessments, model constraints, and forward-looking research agendas. *Technologies, 12*(11), 222. https://doi.org/10.3390/technologies12110222

Harris, S., Hadi, H. J., Ahmad, N., et al. (2025). Multi-domain urdu fake news detection using pre-trained ensemble model. *Scientific Reports, 15*(1), 8705. https://doi.org/10.1038/s41598-025-91054-4

Harris, S., Liu, J., Hadi, H.J., et al.: Ax-to-grind urdu: benchmark dataset for urdu fake news detection. In: 2023 IEEE 22nd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pp. 2440–2447 (2023). DOI: 10.1109/TrustCom60117.2023.00343. IEEE.

Harris, S., Liu, J., Hadi, H. J., et al. (2025). Benchmarking hook and bait urdu news dataset for domain-agnostic and multilingual fake news detection using large language models. *Scientific Reports, 15*(1), 15553. https://doi.org/10.1038/s41598-025-98271-x

Huang, K., Li, X., & Uddin, S. (2026). Enhancing fake news detection through fact-augmented llm generation with co-attention. *Journal of Intelligent Information Systems, 64*(1), 425–443. https://doi.org/10.1007/s10844-025-01007-6

Jin, Z., Cao, J., Guo, H., et al.: Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: Proceedings of the 25th ACM International Conference on Multimedia, pp. 795–816 (2017). DOI: 10.1145/3123266.3123454.

Jiang, D., Liu, Y., Liu, S., et al.: From clip to dino: Visual encoders shout in multi-modal large language models. arXiv:2310.08825 (2023) 10.48550/arXiv. 2310.08825.

Jing, J., Wu, H., Sun, J., et al. (2023). Multimodal fake news detection via progressive fusion networks. *Information Processing & Management, 60*(1), Article 103120. https://doi.org/10.1016/j.ipm.2022.103120

Khattar, D., Goud, J.S., Gupta, M., et al.: Mvae: Multimodal variational autoencoder for fake news detection. In: The World Wide Web Conference, pp. 2915–2921 (2019). DOI: 10.1145/3308558.3313552.

Kaliyar, R. K., Goswami, A., & Narang, P. (2021). Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia Tools and Applications, 80*(8), 11765–11788. https://doi.org/10.1007/s11042-020-10183-2

Kalraa, S., Vermaa, P., Sharma, Y., et al.: Ensembling of various transformer based models for the fake news detection task in the urdu language. In: FIRE (Working Notes), pp. 1175–1181 (2021). https://api.semanticscholar.org/CorpusID:251019957.

Kalra, R., Wu, Z., Gulley, A., et al.: Hypa-rag: A hybrid parameter adaptive retrieval-augmented generation system for ai legal and policy applications. In: Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (customnlp4u), pp. 237–256 (2024). 10.48550/arXiv. 2409.09046.

Lina, N., Fua, S., & Jianga, S. (2020). Fake news detection in the urdu language using charcnn-roberta. *Health, 100*, 100.

Ling, Z., Guo, Z., Huang, Y., et al.: Mmkb-rag: A multi-modal knowledge-based retrieval-augmented generation framework. arXiv:2504.10074 (2025) 10.48550/arXiv. 2504.10074.

Liu, Y., Han, T., Ma, S., et al. (2023). Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-radiology, 1*(2), Article 100017. https://doi.org/10.1016/j.metrad.2023.100017

Li, J., Li, D., Savarese, S., et al.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning, pp. 19730–19742 (2023). PMLR.

Liu, H., Li, C., Wu, Q., et al.: Visual instruction tuning. Advances in Neural Information Processing Systems 36, 34892–34916 (2023) 10.48550/arXiv. 2304.08485.

LekshmiAmmal, H. R., & Madasamy, A. K. (2025). A reasoning based explainable multimodal fake news detection for low resource language using large language models and transformers. *Journal of Big Data, 12*(1), 46. https://doi.org/10.1186/s40537-025-01093-x

Li, X., Qiao, J., Yin, S., et al. (2025). A survey of multimodal fake news detection: a cross-modal interaction perspective. *IEEE Transactions on Emerging Topics in Computational Intelligence*. https://doi.org/10.1109/TETCI.2025.3543389

Liu, Y., Zhu, J., Liu, X., Tang, H., Zhang, Y., Zhang, K., Zhou, X., Chen, E.: Detect, investigate, judge and determine: A knowledge-guided framework for few-shot fake news detection. arXiv:2407.08952 (2024).

Nezafat, M.V., Samet, S.: Fake news detection with retrieval augmented generative artificial intelligence. In: 2024 2nd International Conference on Foundation and Large Language Models (FLLM), pp. 160–167 (2024). DOI: 10.1109/FLLM63129.2024.10852474. IEEE.

Nie, S., Zeng, Z.: Towards real-world multimodal propagation networks: Multimodal uncertainty graph contrastive learning for fake news detection. Journal of Intelligent Information Systems, 1–21 (2025) 10.1007/s10844-025-00987-9.

Qin, H., Ding, Y., Zhang, X., et al. (2023a). Diverse sample generation: Pushing the limit of generative data-free quantization. *IEEE Transactions on Pattern Analysis and Machine Intelligence,45*(10), 11689–11706. https://doi.org/10.1109/TPAMI.2023.3272925

Qin, H., Ke, L., Ma, X., et al. (2023b). Bimatting: Efficient video matting via binarization. *Advances in Neural Information Processing Systems,36*, 43307–43321.

Qin, H., Zhang, M., Ding, Y., et al.: Bibench: Benchmarking and analyzing network binarization. In: International Conference on Machine Learning, pp. 28351–28388 (2023c). PMLR.

Qin, H., Zhang, Y., Ding, Y., et al.: QuantSR: Accurate low-bit quantization for efficient image super-resolution. In: Thirty-seventh Conference on Neural Information Processing Systems (2023d). https://openreview.net/forum?id=3gamyee9Yh.

Ren, R., Ma, J., Zheng, Z.: Large language model for interpreting research policy using adaptive two-stage retrieval augmented fine-tuning method. Expert Systems with Applications 278, 127330 (2025) 10.1016/j.eswa.2025.127330.

Ramya, G., Veda Yasaswani, S., Harshitha, P., et al.: Fake news detection using large language models. In: International Conference on Advanced Network Technologies and Intelligent Computing, pp. 124–137 (2024). DOI: https://doi.org/10.1007/978-3-031-83793-7_9.Springer.

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

Sormeily, A., Dadkhah, S., Zhang, X., et al. (2024). Mefand: A multimodel framework for early fake news detection. *IEEE Transactions on Computational Social Systems, 11*(4), 5337–5353. https://doi.org/10.1109/TCSS.2024.3355300

Singhal, S., Kabra, A., Sharma, M., et al.: Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 13915–13916 (2020). DOI: https://doi.org/10.1609/aaai.v34i10.7230.

Suryavardan, S., Mishra, S., Chakraborty, M., et al.: Findings of factify 2: multimodal fake news detection. arXiv:2307.10475 (2023) 10.48550/arXiv. 2307.10475.

Song, C., Ning, N., Zhang, Y., et al. (2021). A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Information Processing & Management, 58*(1), Article 102437. https://doi.org/10.1016/j.ipm.2020.102437

Singhal, S., Shah, R.R., Chakraborty, T., et al.: Spotfake: A multi-modal framework for fake news detection. In: 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), pp. 39–47 (2019). DOI: 10.1109/BigMM.2019.00-44. IEEE.

Singh, P., Srivastava, R., Rana, K., et al.: Semi-fnd: Stacked ensemble based multimodal inferencing framework for faster fake news detection. Expert Systems with Applications 215, 119302 (2023) 10.1016/j.eswa.2022.119302.

Touvron, H., Lavril, T., Izacard, G., et al.: Llama: Open and efficient foundation language models. arXiv:2302.13971 (2023) 10.48550/arXiv. 2302.13971.

Tahmasebi, S., Müller-Budack, E., Ewerth, R.: Multimodal misinformation detection using large vision-language models. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, pp. 2189–2199 (2024). DOI: 10.1145/3627673.3679826.

Turaga, V.S.P., Namin, A.S.: An information reliability framework for detecting misinformation based on large language models. In: 2024 IEEE International Conference on Big Data (BigData), pp. 3599–3608 (2024). DOI: 10.1109/BigData62323.2024.10826052. IEEE.

Thaokar, C.B., Rathod, M., Ahmed, S., et al.: A multi-linguistic fake news detector on hindi, marathi and telugu. In: 2022 OITS International Conference on Information Technology (OCIT), pp. 324–329 (2022). DOI: 10.1109/OCIT56763.2022.00068. IEEE.

Tufchi, S., Yadav, A., & Ahmed, T. (2023). A comprehensive survey of multimodal fake news detection techniques: advances, challenges, and opportunities. *International Journal of Multimedia Information Retrieval, 12*(2), 28. https://doi.org/10.1007/s13735-023-00296-3

Uppada, S.K., Patel, P., B, S.: An image and text-based multimodal model for detecting fake news in osn's. Journal of Intelligent Information Systems 61(2), 367–393 (2023) 10.1007/s10844-022-00764-y.

Wang, W., Ding, L., Shen, L., et al.: Wisdom: Improving multimodal sentiment analysis by fusing contextual world knowledge. In: Proceedings of the 32nd ACM International Conference on Multimedia, pp. 2282–2291 (2024). DOI: 10.1145/3664647.3681403.

Wu, L., Long, Y., Gao, C., et al.: Mfir: Multimodal fusion and inconsistency reasoning for explainable fake news detection. Information Fusion 100, 101944 (2023) 10.1016/j.inffus.2023.101944.

Wang, Y., Ma, F., Jin, Z., et al.: Eann: Event adversarial neural networks for multi-modal fake news detection. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 849–857 (2018). DOI: https://doi.org/10.1145/3219819.3219903.

Wang, J., Mao, H., & Li, H. (2022). Fmfn: Fine-grained multimodal fusion networks for fake news detection. *Applied Sciences, 12*(3), 1093. https://doi.org/10.3390/app12031093

Wang, Y., Wang, S., Cheng, Q., et al.: Visuothink: Empowering lvlm reasoning with multimodal tree search. arXiv:2504.09130 (2025) 10.48550/arXiv. 2504.09130.

Wang, J., Zhu, Z., Liu, C., et al. (2024). Llm-enhanced multimodal detection of fake news. *PlOS One, 19*(10), 0312240. https://doi.org/10.1371/journal.pone.0312240

Wang, W., Zheng, V. W., Yu, H., et al. (2019). A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST), 10*(2), 1–37. https://doi.org/10.1145/3293318

Wu, Y., Zhan, P., Zhang, Y., et al.: Multimodal fusion with co-attention networks for fake news detection. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 2560–2569 (2021). 10.18653/v1/2021.findings-acl.226.

Xue, L., Constant, N., Roberts, A., et al.: mt5: A massively multilingual pre-trained text-to-text transformer. arXiv:2010.11934 (2020) 10.48550/arXiv. 2010.11934.

Xu, P., Shao, W., Zhang, K., et al.: Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024) 10.48550/arXiv. 2306.09265.

Ying, Q., Hu, X., Zhou, Y., et al.: Bootstrapping multi-view representations for fake news detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 5384–5392 (2023). DOI: 10.1609/aaai.v37i4.25670.

Yao, B.M., Shah, A., Sun, L., et al.: End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2733–2743 (2023). DOI: 10.1145/3539618.3591879.

Zhong, Q., Ding, L., Liu, J., et al.: Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. arXiv:2302.10198 (2023) 10.48550/arXiv. 2302.10198.

Zhang, X., Dadkhah, S., Weismann, A. G., et al. (2023). Multimodal fake news analysis based on image-text similarity. *IEEE Transactions on Computational Social Systems, 11*(1), 959–972. https://doi.org/10.1109/TCSS.2023.3244068

Authors and Affiliations

Sheetal Harris[1] · Vinh Thong Ta[2,3] · Marcello Trovati[4] · Ghada Nakhla[1] · Faiza Latif[5] · Ioannis Korkontzelos[1]

✉ Sheetal Harris
Sheetal.Harris@edgehill.ac.uk

✉ Ioannis Korkontzelos
Yannis.Korkontzelos@edgehill.ac.uk

Vinh Thong Ta
Vinh.Ta@cranfield.ac.uk

Marcello Trovati
MTrovati@lancashire.ac.uk

Ghada Nakhla
Ghada.Nakhla@edgehill.ac.uk

Faiza Latif
faiza.prad.scs@pu.edu.pk

[1] Department of Computer Science, Edge Hill University, St Helens Road, Ormskirk L39 4QP, UK

[2] Centre for Defence and Security Management and Informatics, Cranfield University, College Road, Cranfield MK43 0AL, UK

[3] The Defence Academy of the United Kingdom, Shrivenham, Swindon SN6 8LA, UK

[4] Business School, University of Lancashire, Victoria Street, Preston PR1 2HE, UK

[5] Department of Public Relations and Advertising, University of the Punjab, Quaid-e-Azam Campus, Lahore 54590, Pakistan