The Indonesian Journal of Computer Science
www.ijcs.net
Volume 15, Issue 3, June 2026
https://doi.org/10.33022/ijcs.v15i3.5129

Misinformation Detection in Low-Resource Languages and Health Domains: *Review and Evaluation Framework*

Bassey Isong[1], Rose Linah[2]
bassey.isong@nwu.ac.za[1], 32289340@mynwu.ac.za[2]
[1,2]Computer Science Department, North-West University, Mafikeng, South Africa

Article Information

Received : 16 Apr 2026
Revised : 17 Jun 2026
Accepted : 27 Jun 2026

Keywords

Misinformation detection, LRLs, Health misinformation, Explainable AI, Multimodal detection.

Abstract

Misinformation on digital platforms harms public health decisions, electoral processes, and institutional trust. Its detection in low-resource languages (LRLs) remains structurally neglected. No prior survey applies a scoring framework to assess study quality or enable cross-study comparison. This review examines 51 peer-reviewed studies published between 2021 and 2026, following the PRISMA 2020 protocol across four dimensions: detection methodology, dataset coverage, health-domain adaptation, and explainable AI (XAI) integration, and proposes the LRL misinformation evaluation framework (LRLM-EF). The five-criterion evaluation framework was applied retrospectively to all reviewed studies. The findings reveal that dataset construction is the dominant research activity, with new corpora built for Amharic, Bangla, Bengali, Luganda, Sepedi, Sesotho, Xitsonga, isiZulu, Kurdish Sorani, and several Arabic dialects. Transformer-based models outperform classical and deep learning baselines in most settings; classical classifiers achieve comparable results where annotated data is scarce. Health-domain coverage is narrow. Research mostly concentrates on COVID-19 and vaccine misinformation, while HIV, malaria, and reproductive health appear in no reviewed studies. Multimodal fusion improves detection in all five studies where it was tested, yet audio-based detection in any LRL setting is absent. XAI is applied in a few studies, exclusively through post-hoc LIME, with no study evaluating its effect on user decisions. LRLM-EF scoring reveals that most studies address fewer than half the framework criteria, with adversarial evaluation and standardised reporting as the weakest dimensions. However, two contradictions exist in the evidence. Classical retrieval outperforms neural similarity on rare-terminology datasets, and augmentation volume shows no reliable accuracy gain, which further exposes the absence of a shared benchmarking standard.

Attribution-ShareAlike 4.0 International License

A. Introduction

The swift rise of online social networks has fundamentally altered how information circulates globally, providing a low-cost, largely unregulated channel for communication [1], [2], [3]. This shift has consequences. The same properties that enable rapid information sharing have made these platforms vectors for misinformation [4], [5], [6]. Misinformation, false or misleading content unsupported by scientific evidence, takes the form of fake news, disinformation, rumours, and satire [1], [5], [7]. The World Economic Forum ranks it among the top ten global risks, citing its corrosive effects on public trust, democratic institutions, and social cohesion [4], [5], [8], [9]. The consequences are particularly severe in healthcare. During the COVID-19 pandemic, an overabundance of both accurate and false information, termed an "infodemic" by the World Health Organization, drove vaccine hesitancy, promoted dangerous remedies, and caused large-scale societal disruption [4], [8-13]. The problem extends well beyond the pandemic. Inaccurate advice on diabetes management, maternal health, and neonatal care from unqualified sources has been linked directly to medical complications and death [14-16]. Politically, misinformation has altered electoral outcomes, with the 2016 US presidential election widely cited as an early case [17-19].

Before the digital age, false information spread through word-of-mouth and print at rates that allowed some degree of correction. That is no longer the case. Social media platforms such as Facebook and X (formerly Twitter) allow unverified content to go viral without editorial scrutiny, and the volume and velocity of online content have made manual fact-checking by journalists and domain experts unworkable [17], [18], [20-22]. This necessitated the need for automated detection. Early systems used traditional machine learning (ML), but the field has since moved toward deep learning (DL) architectures, particularly transformer-based models such as BERT, which capture semantic relationships and linguistic patterns that classical methods miss [17], [19], [22], [23].

Today, detection has grown more complex. Recent frameworks incorporate visual elements and platform metadata from sources including TikTok and YouTube, and hybrid approaches combine neural networks with fuzzy logic or probabilistic modelling to make outputs more interpretable [24-26]. Metaheuristic optimisation and domain-specific transformer models have also entered the field [25], [26]. However, three problems remain unresolved. Annotated datasets and detection tools for low-resource languages (LRLs) are scarce. Arabic dialects, Amharic, Kurdish, Bangla, and southern African languages, including Sepedi, Sesotho, Tshivenda, and Xitsonga, are largely absent from existing benchmarks [4-6], [15-19], [22]. Class imbalance in training data and modality gaps across audio, image, and video content limit how well models generalise. Beyond performance, the spread of social bots and adversarial text manipulation requires systems that can explain their decisions, not just produce them, if they are to earn and sustain human trust [1], [4], [9].

This review addresses these gaps through four contributions. First, it presents a review of 51 peer-reviewed studies, combining trends in dataset development, modelling techniques, and explainability integration with a particular focus on LRLs and health-related domains. Second, it analyses ML, DL, transformer-based, and hybrid approaches employed across the reviewed studies. Third, it assesses publicly available and newly created datasets across multiple languages and modalities. Fourth, it examines the integration of explainable AI methods, specifically LIME, SHAP, and Eli5, within misinformation detection (MID) frameworks and identifies important research directions necessary to advance the field.

The rest of the paper is structured as follows: Section B presents the related works, Section C presents the methodology used, and Section D presents the findings of the review. Section E discusses the findings and the future research opportunities, while

Sections F and G present the proposed evaluation framework and conclude the paper, respectively.

B. Related Works

B.1. Misinformation detection overview and approaches

Currently, there is no universally accepted definition of fake news that exists. The term has been applied to contexts ranging from sensational newspaper reports in the 1890s to viral disinformation on contemporary social media [18]. The literature converges on a taxonomy organised by intent and authenticity [21], [27]. Three core types are distinguished: deliberately false disinformation; misinformation, which is false but unintentional; and malinformation, which uses genuine information to cause harm [21], [22]. Specific manifestations include rumours, clickbait, fabricated content designed to mimic credible news formats, satire, and conspiracy theories that proliferate in social media echo chambers [18], [20], [21], [27-29].

Dataset coverage is uneven. Most research targets English-language textual content, with 41 public and 58 private datasets identified for English alone [18]. Languages such as Arabic, Urdu, and Persian each have only one public dataset. Standard benchmarks: LIAR, ISOT, and FakeNewsNet are the most widely cited [17], [19], [21], [23]. Domain-specific datasets emerged during the COVID-19 pandemic, such as CoAID, FakeHealth, and MMCovid19, which were built to address health misinformation [17-19], [29], [30]. Despite this growth, multimodal datasets remain scarce, and audio and video corpora are largely absent from LRL settings. Detection approaches fall into two broad paradigms. Content-based methods analyse lexical, syntactic, and stylistic features internal to a news item, including headline-body similarity and sentiment polarity [19], [21], [23], [27], [29]. Social context and propagation-based methods examine user engagement patterns, propagation network structures, and account metadata to identify behaviour consistent with automated bots [18], [19], [21], [23]. These paradigms are not mutually exclusive; recent systems combine both.

Early detection systems used SVM, RF, and Naïve Bayes classifiers for linguistic feature analysis [9], [16], [19], [22], [23], [27]. RF remained competitive in scenarios with mature feature engineering or limited computational resources, reporting accuracy up to 96-97% on lexical and metadata features [6], [9], [16], [29], [30]. DL addressed the limitations of classical methods directly. CNNs extract local features such as n-gram patterns, while LSTM and BiLSTM networks capture sequential and temporal dependencies [6],[9],[10],[17],[19]. CNN-BiLSTM hybrids combine local feature extraction with long-range dependency modelling, reaching accuracy as high as 98.89% on benchmark datasets [31]. Transformer-based architectures now set the performance standard. BERT and RoBERTa model context bidirectionally via self-attention, with accuracy reported as high as 99.9% on standard benchmarks [17],[19], [27], [30]. Domain-specific pre-training yields further gains: CT-BERT, pre-trained on COVID-19 Twitter data with a BiGRU layer, reached an F1-score of 98% [33]. Multilingual variants extend these gains: mBERT supports low-resource transfer [4], [5], [34], AraBERT and MARBERT address Arabic dialects [1], [5], [14], and XLM-RoBERTa reaches F1-scores of 98% on multilingual news classification [1], [8], [34].

Text alone is insufficient for contemporary misinformation. Modern MID frequently exploits images, audio, and video [5],[6],[10],[21],[22],[30],[35]. Multimodal frameworks pair visual encoders with language models, reaching 82-90% accuracy [4], [25]. Platform-specific work extends these methods: on YouTube, using related video metadata improved detection accuracy by up to 9.7% [25]; for TikTok, hybrid frameworks integrating speech-to-text conversion with personality trait analysis generate interpretable reports for human auditors [24]. Incorporating user interaction patterns like retweets, replies,

account age, follower counts, and propagation paths regularly improves detection [1], [5], [7], [16]. Social bots account for approximately 10% of misinformation and adapt their behaviour over time to evade detection [38]. Knowledge graphs and automated fact-checking pipelines generate explicit rationales for detection outcomes, which matters in clinical contexts where model decisions require justification [30],[41]. Post-hoc XAI methods, specifically LIME, SHAP, and Eli5, identify influential words and provide local explanations, making model reasoning accessible to non-expert users [1],[4],[9],[42]. However, XAI adoption is uneven as its use in LRL settings lags considerably behind high-resource settings.

B.2. Challenges in misinformation detection

Research has shown that over 67% of MID studies focus on English [18], [22]. Speakers of African and other LRLs remain substantially underserved. Cross-lingual frameworks such as CrossFake transfer fact-checked data from English to low-resource settings via machine translation and sub-text aggregation, though translation quality remains a bottleneck [12]. Transfer learning and cross-lingual embeddings offer a pathway forward, but the absence of large labelled LRL datasets limits how far these methods generalise [19], [22]. Four additional problems cut across the literature. Class imbalance skews classifiers toward the majority class when training data contains far more legitimate news than misinformation [17], [20]. Generalisation across topics is unresolved: models trained on established datasets frequently fail on emerging misinformation [17], [19]. Adversarial vulnerability means detection systems are evaded by creators who continuously adapt their tactics; XAI systems themselves can also be manipulated through adversarial attacks [17], [19], [20]. Institutional bias in training data sourced from politically affiliated fact-checking agencies without independent auditing may embed partiality into detection models [19]. These problems indicate a field where benchmark performance has outpaced deployment readiness in diverse operational conditions.

B.3. Related Studies

Several reviews and surveys have examined MID from different perspectives. None addresses the intersection of LRLs, health domains, multimodality, and explainability within a single framework. Alnabhan and Branco [17] traced CNN and LSTM architectures across 176 studies, framing transfer learning as a response to data scarcity and class imbalance. Hussain et al. [18] drew on 310 articles and 271 datasets to document the shift from TF-IDF features to transformer-based models, noting that multimodal data and languages outside English remain largely unexplored. Tata et al. [27] found that CNNs and transformer models outperform SVMs across their reviewed corpus. Alkomah and Sheldon [23] reviewed supervised, unsupervised, and ensemble approaches, finding that adversarial resilience and ethical standards remain unresolved. Bodaghi et al. [21] distinguished misinformation from deliberate disinformation. Bashaddadh et al. [19] found transformer-based models ahead of traditional classifiers and flagged XAI and few-shot learning as unresolved priorities.

Health and language-specific reviews add further context. Feng et al. [30] analysed 100 studies spanning ML, DL, knowledge graphs, multimodal fusion, and large language models, concluding that interdisciplinary collaboration is a prerequisite for clinically relevant detection. Similarly, Schlicht et al. [29] proposed a taxonomy for health misinformation and showed that pandemic rumours differ linguistically and emotionally from other medical falsehoods; data scarcity and the absence of XAI were identified as the two primary gaps. Rananga et al. [22] also applied PRISMA to 31 studies and found that two-thirds of MID research focuses on English, with Chinese as the only other language

with substantial representation, calling for culturally grounded datasets to reduce LRL community exposure to undetected misinformation.

These reviews establish a solid foundation in English-language, text-based detection but leave the overlap between LRL contexts and health-domain misinformation unexamined, and treat explainability as a peripheral concern rather than an analytical dimension. This review fills that gap, analysing 51 peer-reviewed studies across four dimensions: detection methodology, dataset coverage across languages and modalities, health-domain adaptation, and XAI integration.

C. Research Method

The review followed the PRISMA 2020 framework [46],[47] to ensure transparency and reproducibility throughout study identification, screening, and synthesis. Four research questions guided the process: what trends characterise MID research published between 2021 and 2026, particularly for LRLs; which ML and DL algorithms have been applied and how their use has shifted over time; how detection models are evaluated and which metrics and validation strategies are reported; and what emerging directions are shaping the field. To achieve this, six academic databases were searched: Google Scholar, IEEE Xplore, ResearchGate, ScienceDirect, SpringerLink, and MDPI. The search string combined four central concepts using Boolean operators: ("fake news" OR "misinformation") AND ("detection" OR "classification") AND ("low-resource languages" OR "health misinformation" OR "HIV/AIDS" OR "multimodal" OR "explainable AI"). It was applied to titles, abstracts, and keywords of peer-reviewed papers published between January 2021 and January 2026. The initial search returned 312 records.

Inclusion required studies to be peer-reviewed journal articles or conference papers in English, situated within computer science or NLP, and employing ML or DL methods for misinformation or fake news detection, including work incorporating XAI or interpretability. Studies also had to address multilingual, low-resource, or health-domain contexts and report at least one quantitative metric such as accuracy, F1-score, precision, or recall. Excluded were non-academic works, including blogs, theses, patents, and preprints; non-English papers; studies outside computer science or NLP; papers lacking methodological detail or performance results; and inaccessible full texts. Screening proceeded in three phases. In the identification phase, 312 records were retrieved. After removing 74 duplicates, 238 unique records remained. Titles and abstracts were then examined; 142 records were excluded for covering unrelated domains such as spam filtering or sentiment analysis without a misinformation context. The remaining 96 full texts were assessed for eligibility. Of these, 45 were excluded for lacking empirical evaluation, omitting performance metrics, or presenting purely conceptual work without computational implementation. 51 studies met all criteria and were included for synthesis. Fig. 1 presents the full PRISMA-based workflow.

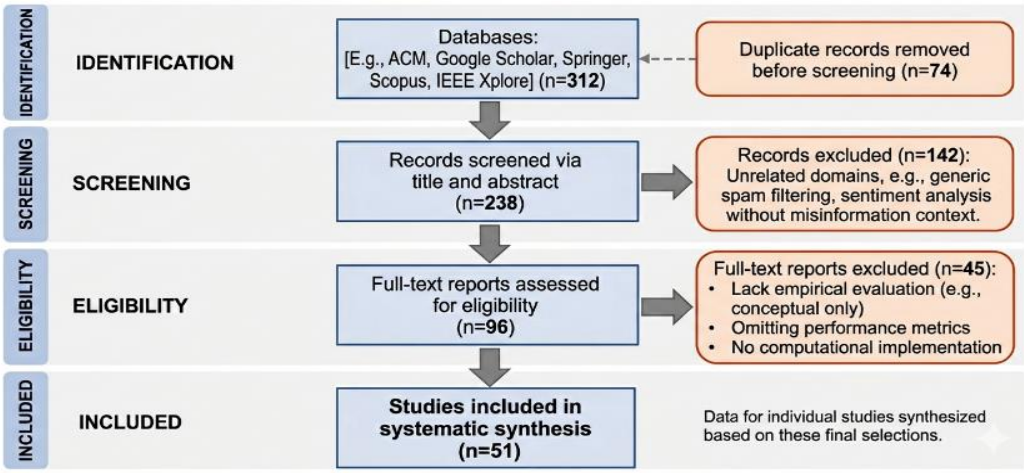

Figure 1. Study Selection Workflow

Data were extracted using a structured spreadsheet template applied consistently across all 51 studies. Several attributes were recorded per study: publication year and venue; misinformation domain (general, political, or health-related); data modality (text, image, video, or multimodal); word embedding strategy; classifier type (ML, DL, transformer, hybrid, graph-based, or ensemble); dataset name and origin; evaluation metrics; validation technique; baseline comparisons; results; language resource level (LRL or HRL); language; explainability methods (LIME, SHAP, Eli5, or other); and reported limitations. Extracted data were analysed quantitatively and qualitatively to identify trends across model types, languages, and datasets. In addition, coded data were synthesised thematically across five dimensions: language focus; dataset characteristics, including size, source, and augmentation strategies; model type; performance metrics; and key contributions covering dataset creation, methodological innovation, and explainability. Findings were cross-checked across multiple studies and triangulated across domains: health, politics, and social media, to confirm consistency. Comparative tables were constructed to map differences and commonalities, supporting a structured combination of results.

D. Result

We organized the 51 reviewed studies into a coherent analytical structure by classifying each study along four dimensions: resource setting, input modality, detection method family, and XAI status. Table 1 summarises the distribution across each dimension. This taxonomy offers a reference map of the field as it currently stands and grounds the LRLM-EF framework proposed in Section 6 by making the field's gaps numerically visible.

Table 1. Distribution of reviewed studies across four dimensions (n = 51).

Dimension	Category	Studies (n)	Notes
Resource context	LRL (low-resource)	31	New corpora; transfer learning from HRL
Resource context	HRL (high-resource)	20	English-dominant benchmark datasets
Input modality	Text only	44 (86%)	Dominant mode across LRL and HRL
Input modality	Text + image	5 (10%)	Multimodal fusion; LRL and HRL

Input modality	Text + audio/video	2 (4%)	Both studies use English or Spanish
Method family	Transformer-based	28	XLM-RoBERTa, mBERT, AraBERT, MARBERT
Method family	Classical ML	14	SVM, RF, LR, NB; competitive under data scarcity
Method family	Deep learning	10	LSTM, CNN, BiGRU
Method family	LLM / hybrid ensemble	8	GPT-3/4, BLOOM, stacking ensembles
XAI status	XAI applied	10	LIME (9), SHAP (1), ELi5 (1) - all post-hoc
XAI status	No XAI	41	Black-box outputs only

The distribution shown in Table 1 reveals three structural patterns. Text-only detection accounts for 86% of the corpus, a concentration that understates the multimodal nature of misinformation in practice. Transformer-based models appear in 55% of studies, yet classical classifiers remain the method of choice in severe data-scarcity settings. This poses a finding that bears on how performance results should be interpreted across studies. XAI adoption is limited to only ten studies, all applying post-hoc methods, with none evaluating the effect of explanations on user decisions.

This section presents the findings from the review conducted. We thematically presented them based on their focus: LRL, health information, multimodal detection, and model evolution. These studies are summarized in Tables 2 to 6, focusing on primary language(s), approximate size, modalities, metrics, findings, and XAI.

D.1. Low-resource language detection

A. Amharic

MID in the Amharic language is discussed here. Gereme et al. [48] addressed the complete absence of NLP infrastructure for Amharic fake news detection by building three purpose-made resources: the General-Purpose Amharic Corpus (GPAC), the AMFTWE word embedding model, and the ETH_FAKE dataset. No XAI was applied. A CNN trained on these assets achieved validation accuracy above 99% with both 300- and 200-dimensional embeddings. Extending the approach to closely related dialects is difficult, given its dependence on data volume, which the resources do not yet provide. Gemeda et al. [1] also constructed the Ethio-Fake framework for Amharic social media misinformation, introducing a 12,000-article annotated dataset and evaluating SVM, Logistic Regression (LR), Multinomial Naïve Bayes, and fine-tuned mBERT. LIME was applied to explain individual classification decisions. Fine-tuned mBERT reached 94% F1, ahead of SVM at 93%. The dataset is small relative to what stable transformer generalisation typically requires. Similarly, Geto et al. [44] built the first multimodal Amharic fake news system, assembling over 23,000 Facebook news stories combining text and images. A late-fusion CNN with attention-based BiLSTM was used; no XAI was applied. The hybrid reached 98% accuracy. Coverage is limited to Facebook, and generalisation to other platforms is untested.

B. South African languages

In this category, Mukwevho et al. [49] produced the first MID resource for Tshivenda, web-crawling social media posts, and training CRF, GRU, and LSTM models without XAI. GRU reached 92.84% accuracy, ahead of LSTM at 92.42% and CRF at 92.58%. The dataset origin is undisclosed, limiting reproducibility. On the same note, Manaswe et al. [3] translated the English COVID-19-FNIR dataset into Sepedi via Google Cloud API and

applied six augmentation strategies: back-translation, synonym replacement, random insertion, deletion, and word swap.

SVM and the Zabantu multilingual RoBERTa model were evaluated; no XAI was used. SVM reached 99% accuracy on fully augmented data; Zabantu performed best at 75% back-translation augmentation. Gaps in Sepedi synonym databases constrain what augmentation can achieve. Also, Mokoena et al. [50] translated CoAID into Sesotho using both Google Translate and Helsinki-NLP Opus-MT, with a native linguist validating a portion of the output. Six ML classifiers were tested without XAI. SVM achieved 91% accuracy on the Helsinki-NLP version, outperforming the Google Translate variant. Uninformative translated tokens persist in the corpus and reduce feature discriminability.

Table 2. Summary of Reviewed Studies

Ref.	Dataset · language · size	Modalities	Model/method	Metrics	XAI	Best result	Resource
[1]	Ethio-Fake corpus · Amharic · 12,000	Text	SVM, LR, NB, mBERT (fine-tuned)	F1, Prec, Rec	LIME	F1: 94% (mBERT)	LRL
[4]	CoAID & MiSoVac · English · not stated	Text, Image, Tab.	DANN	F1, Acc, AUC	LIME	Acc: +22%; AUC: +15% vs base	HRL
[8]	BanFakeNews-2.0 · Bangla · 60,000	Text	BLOOM 560M, Phi-3, Llama 3.2, mBERT	Macro F1	None	Macro F1: 89% (BLOOM)	LRL
[9]	LIAR, FakeNewsNet, BuzzFeed · English · benchmark	Text	BERT + RoBERTa + XLNet + GNN + GCN	Acc, Prec, F1	LIME, SHAP	Acc: 96.2% (RoBERTa)	HRL
[14]	Headline-article pairs · Bangla · 8,500	Text	BERT + XGBoost (voting + boosting ensemble)	Acc, F1, AUC	None	Acc: 98%; F1: 99%	LRL
[36]	MMIFND · Multi-Indic (7 langs) · 28,000	Text, Image	FLAVA, BLIP-2, MuRIL (MMCFND)	Acc, Prec, Rec, F1	None	F1: 99.4%	LRL
[48]	GPAC & ETH_FAKE · Amharic · not stated	Text	FastText + CNN	Acc, F1, Prec	None	Acc: >99%	LRL
[49]	Social posts · Tshivenda · not stated	Text	CRF, GRU, LSTM	Acc	None	Acc: 92.84% (GRU)	LRL
[53]	AraEC & AraCoV · Arabic · not stated	Text	AraBERT, mBERT, XLM-R + CNN/LSTM (late fusion)	Acc, Macro F1	None	Acc: 98.89%	LRL
[61]	Twitter vaccine · PT/ID/NG English · not stated	Text	XLM-RoBERTa + GPT-3 augmentation	Macro F1	None	F1: 97.8% (XLM-R+_AUG)	LRL
[66]	AMI dataset · English + Italian · not stated	Text	CNN-LSTM + FastText; FLAN-T5 (incremental)	Acc, F1	LIME	Acc: 95% (CNN-LSTM)	HRL

**LRL = low-resource language; HRL = high-resource language.

Baloyi et al. [51] applied synonym replacement and MarianMT augmentation to a translated COVID-19 Xitsonga dataset, then fine-tuned mBERT against an RF baseline. No XAI was used. mBERT reached 98.7% accuracy, 84.6% precision, and 88% F1; RF reached only 33.3% precision. The dataset is small, which constrains generalisation. Sethosa et al. [52] equally translated an English Twitter hate speech dataset into isiZulu and refined it with native speakers. SVM, LR, and RF were trained using TF-IDF, without XAI. SVM achieved 82.49% accuracy and 83.02% precision. Publicly available isiZulu data is scarce, and translated text cannot fully capture idiomatic usage. Finally, Wet and Marivate [43] tested whether South African English disinformation carries stylistic features distinct from US English. They paired local articles with Media Monitoring Africa verified content and trained LR and LSTM models, applying LIME to surface cues such as repetitive phrasing and exaggerated sentiment. Combining both datasets reached 90.6% accuracy, against 80–

Table 3. Summary of Reviewed Studies

Ref.	Dataset · language · size	Modalities	Model/method	Metrics	XAI	Best result	Resource
[5]	YouTube comments · Algerian Arabic · ~1,500	Text	SetFit + MARBERTv2, DziriBERT (few-shot)	Acc, F1, Prec	None	F1: 71% (DziriBERT)	LRL
[15]	KDFND · Kurdish (Sorani) · 100,000+	Text	SVM, RF, CNN	Acc, F1	None	Acc: 91.6%; F1: 95.47% (CNN)	LRL
[34]	Dravidian_Fake · 4 Dravidian langs · ~26,000	Text	mBERT, XLM-RoBERTa (zero/cross/multilingual)	Acc, F1	None	Acc: 93.31%; F1: 93.29% (XLM-R)	LRL
[40]	Mpox & Polio (Twitter) · English · not stated	Text	CrowdAdapt (GKE + DRL + CKU)	Acc, Prec, F1	None	F1: +9.1% vs DETERRENT (Mpox)	HRL
[42]	Arabic Fake News Dataset · Arabic · not stated	Text	ELMo + CNN-LSTM (voting ensemble)	Acc, F1, Rec	LIME	Acc: 98.42%; F1: 98.93%	LRL
[55]	ANS Corpus & AraNews · Arabic · ~2,500	Text	MARBERTv2 (multitask fine-tuned)	F1, IAA (Fleiss κ)	None	F1: 94.12% (AraNews)	LRL
[57]	UST · Urdu · not stated	Text	SVM, RF, LR, Naïve Bayes	F1	None	F1: 84% (SVM)	LRL
[58]	YouTube audio · English + Kiswahili · not stated	Text, Audio, Video	RF, XGBoost, Bi-LSTM, BERT, GPT-J-6B	Acc, Prec, Rec, F1	LIME	Acc: 98% (RF, EN); F1: 94% (SW)	LRL
[59]	Mendeley COVID · English · not stated	Text	CNN + GloVe (stratified k-fold)	Acc, F1, AUC	None	Acc: 96.19%	HRL
[63]	Crowdsourced health · English · not stated	Text	LR	Acc, Prec	ELi5	Train acc: 93%; Prec: 83%	HRL

**LRL = low-resource language; HRL = high-resource language

84% on single-domain training. The system sees stylistic surface features only; broader discourse context is outside its scope.

C. Arabic and its dialects

For the Arabic language, Hossaine et al. [53] built AraCovTexFinder to identify and classify COVID-19-related Arabic texts. They created the AraEC and AraCoV datasets and integrated AraBERT, BERT-Arabic, mBERT, XLM-RoBERTa, mDeBERTa-V3, and mDistilBERT with CNN, LSTM, VDCNN, and BiLSTM through weighted late fusion. No XAI was applied. The system achieved 98.89% accuracy across 18 baseline comparisons. The dataset size was constrained by robots.txt crawling restrictions. In the same vein, Aboulola

Table 4. Summary of Reviewed Studies

Ref.	Dataset · language · size	Modalities	Model/method	Metrics	XAI	Best result	Resource
[6]	Hindi-English curated · Hindi+English · ~30,000	Text	BiLSTM, 1DCNN-BiLSTM	Acc, Prec, F1	None	Acc: 92.36% (CNN-BiLSTM)	LRL
[10]	MM-COVID · Multilingual · not stated	Text, Image	BERT-BiLSTM (text) + GNN-CNN (image)	Acc, F1, AUC	None	Acc: 99.93% (text only)	HRL
[16]	Zibizheng Ba autism forum · Chinese · ~5,000	Text	ML classifiers + behavioural features	F1, Prec	None	~85% misinfo detected (RF)	HRL
[24]	TikTok survey · English + Spanish · not stated	Text, Audio, Video	Fuzzy + DL hybrid (GLMP)	Prec, Rec, F1	GLMP reports	Multimodal F1: 95%	HRL
[38]	COVID-19 tweets · English · 10.22M	Text	Stacking ensemble (SVM, RF, DT, KNN, LR)	Prec, Rec, FPR	None	Prec: 83%; Rec: 96%	HRL
[39]	Indic-COVID-19 · Hindi + Bengali · not stated	Text	BERT + mBERT + user metadata; zero-shot	Prec, Rec, F1	None	Zero-shot F1: 81% (Hindi)	LRL
[54]	Twitter Arabic COVID · Arabic · not stated	Text	AraBERT, MARBERT (+ oversampling)	F1, Acc	None	Acc: 97%; F1: 96% (MARBERT)	LRL
[56]	BFNC · Bengali · 5,000+	Text	XLM-RoBERTa, Bangla-BERT, mBERT, LR	Acc, F1	None	F1: 98% (XLM-R)	LRL
[62]	Arabic diabetes social media · Arabic · 10,000+	Text	LSTM, DistilBERT, GPT-4	Acc, Prec	None	Acc: 96% (DistilBERT)	LRL
[64]	Bilingual health corpora · English + Malay · not stated	Text	DT, RF, SVM (lexical features)	Acc, F1, AUC	None	Acc: 95%; F1: 92.8% (RF, Malay)	LRL

**LRL = low-resource language; HRL = high-resource language

and Umer [42] targeted Arabic fake news classification using ELMo contextual embeddings with a CNN-LSTM voting ensemble, applying LIME to visualise word-level influences. The system achieved 98.42% accuracy, 98.54% precision, 99.5% recall, and 98.93% F1. Computational cost is the primary deployment barrier.

Bahurmuz et al. [54] applied AraBERT and MARBERT with oversampling to Arabic COVID-19 Twitter rumour detection, without XAI. MARBERT reached 97% accuracy and 96% F1. The approach does not transfer well to low-resource Arabic dialects beyond Modern Standard Arabic and Gulf variants. Abdedaiem [5] also used the SetFit few-shot framework to detect fake news in the Algerian Arabic dialect, testing MARBERTv2, DziriBERT, DarijaBERT, ARBERT, CAMeLBERT, AraBART, and AraBERTv2 without XAI. Dialect-specific models outperformed MSA models; DziriBERT reached 71% F1 on the Algerian dialect. Only 1,500 samples were used, which constrains what can be concluded. Similarly, Shehata et al. [55] developed ArabFake as a multitask MARBERTv2 framework performing fake news detection, misinformation categorisation, and risk assessment simultaneously. Valence scoring was incorporated as an interpretability mechanism. ArabFake achieved 94.12% F1 on AraNews and 69% on the ANS Corpus. Performance varies substantially across corpora, indicating sensitivity to domain and style.

D. Indic and South Asian languages

From the Asian perspective, Shibu et al. [8] benchmarked large language models on the newly constructed BanFakeNews-2.0, a 60,000-item Bangla dataset across 13 categories. BLOOM 560M, Phi-3, Llama 3.2, mBERT-uncased, and the C3-Gram baseline were evaluated without XAI. BLOOM 560M reached 89% macro F1, against C3-Gram at 77%. The dataset skews toward real news, biasing classifiers toward the majority class. Equally, Sultana and Nishino [14] combined BERT with voting regressors and XGBoost boosting for Bangla fake news detection, integrating sentiment, emotion, hate speech, irony, and grammar features. No XAI was used. The model reached 98% accuracy, 99% recall, and 92% AUC. Source verification is absent, limiting applicability to outlets without structured metadata. Rahman et al. [56] built the Bengali Fake News Corpus, over 5,000 manually annotated items, and proposed the FaND-X framework. XLM-RoBERTa, Bangla-BERT, mBERT, and LR were evaluated; error analysis served as an interpretability proxy rather than a formal XAI method. XLM-RoBERTa reached 98% F1; LR reached 89%. Satire misclassified as misinformation is identified as a recurring failure.

Borgohain et al. [34] also introduced Dravidian_Fake, covering approximately 26,000 articles across Telugu, Tamil, Kannada, and Malayalam, and compared zero-shot, cross-lingual, and multilingual fine-tuning of mBERT and XLM-RoBERTa without XAI. XLM-RoBERTa reached 93.31% average accuracy and 93.29% F1 in multilingual transfer learning. Models overfit on training language distributions, reducing cross-lingual stability.

In other studies, Nandgaonkar et al. [6] built a Hindi-English dataset of nearly 30,000 articles and trained BiLSTM and 1DCNN-BiLSTM models without XAI. The hybrid reached 92.36% accuracy, marginally ahead of BiLSTM at 92.23%. Hindi NLP tooling gaps constrain preprocessing quality. Kar et al. [39] introduced the Indic-COVID-19 Fake Tweet Dataset for Hindi and Bengali and combined BERT and mBERT with user metadata features: authority scores and network indicators. No XAI was used. Zero-shot accuracy reached 81% for Hindi and 78% for Bengali; the English model achieved 89% F1. However, the dataset size limits the robustness of the zero-shot transfer findings. In parallel, Bansal et al. [36] developed MMIFND, a multimodal dataset covering seven Indic languages, and the MMCFND system fusing text, images, and captions through FLAVA and BLIP-2. No XAI was applied. The system reached 99.6% precision, 99.8% recall, and 99.4% F1. Ablation confirmed that removing text dropped F1 by 9.3% for fake samples. Dynamic image loading at inference is an engineering constraint.

E. Other low-resource languages

Studies focusing on other languages are discussed here. Khan et al. [57] built the Urdu Sarcastic Tweets dataset and evaluated SVM, RF, LR, and Multinomial Naïve Bayes using N-grams and emoji-handling features, without XAI. SVM reached 84% F1, against the lexicon-based Tanz-Indicator at 32.3%. These results need larger corpora before they can be treated as definitive. Salh and Nabi [15] also introduced the Kurdish Sorani Fake News Dataset, over 100,000 articles, and evaluated SVM, RF, and CNN with word embeddings and TF-IDF, without XAI. CNN reached 91.6% accuracy and 95.47% F1, against RF at 73.37%. Default hyperparameters were used, which may understate achievable performance.

Nabende et al. [2] produced the first Luganda misinformation dataset from Facebook and Twitter posts, evaluating traditional and ensemble ML classifiers in Weka without XAI. DMNBtext achieved 78.19% accuracy and 76.9% F1. Code-switching annotation is an unresolved challenge in the corpus. Similarly, Imbwaga et al. [58] constructed a hate speech pipeline for English and Kiswahili using audio from YouTube. FFmpeg extracted audio, Whisper AI transcribed it, and SVM, RF, XGBoost, Bi-LSTM, and GPT-J-6B were evaluated with LIME applied to identify words driving hateful classifications. RF achieved 98% accuracy on English and 94% F1 on Kiswahili. Tonal and prosodic features are not captured by the text-only classification stage.

D.2. Health domain misinformation

A. COVID-19

In this category, Akhter et al. [59] detected COVID-19 misinformation using a CNN with GloVe embeddings, grid search for hyperparameter tuning, and stratified k-fold cross-validation to handle dataset imbalance. No XAI was applied. The CNN achieved 96.19% mean accuracy and 95% F1. Training data skews toward legitimate content, constraining recall on misinformation samples. Hayawi et al. [11] also produced the ANTi-Vax dataset with over 15,000 expert-labeled vaccine tweets, and tested BERT, XGBoost, and LSTM against it without XAI. BERT reached 98% accuracy and 98% F1. Bot amplification behaviour falls outside the study's scope. Similarly, Alghamdi et al. [33] compared generic BERT against CT-BERT, pre-trained on COVID-Twitter data, adding BiGRU and CNN downstream layers with fine-tuning and no XAI. CT-BERT with BiGRU reached 98% F1, confirming the advantage of domain-specific pre-training over general transformers. The training dataset is small, which may limit transfer beyond the study period. While Joshi et al. [4] applied Domain Adversarial Neural Networks to CoAID and MiSoVac, incorporating LIME for prediction-level explanations. Twitter accuracy improved by 22% and AUC by 15% over non-adversarial baselines. DANN's internal reasoning remains opaque beyond what LIME exposes at the instance level.

Weinzierl and Harabagiu [41] reframed vaccine MID as graph link prediction on the COVAXLIES dataset. Tweets were represented as nodes in a Misinformation Knowledge Graph using COVID-Twitter-BERT-v2 embeddings; no post-hoc XAI was applied. TMD-GLP achieved 84.3% F1, outperforming TMD-BC-BERT at 76.8% and TMD-BC-LSTM at 54.9%. Modelling complex relational structures between misinformation targets is computationally demanding. The knowledge graph structure also functions as a social context model, connecting tweets through shared misinformation targets rather than user propagation paths, a conceptually distinct form of context-aware detection. Also, Xia et al. [60] tracked COVID-19 misinformation across its lifecycle using CNN-BiLSTM-AM with multi-head attention and an outlier knowledge management framework, without XAI. The model achieved 93.6% accuracy; fake news clustered around fear-driven themes, including virus origins and vaccine safety. Added complexity introduced overfitting at finer granularity levels. Akhtar et al. [38] cross-referenced tweets against verified fact-

checked statements and built a stacking ensemble of SVM, RF, Decision Tree (DT), KNN, and LR with an LR meta-learner. Semantic matching served as an interpretability mechanism. The ensemble reached 83% precision and 96% recall. Bot-generated tweets showed approximately 3% engagement versus under 1% for human tweets. Manual verification does not scale.

Du et al. [12] built CrossFake to detect Chinese COVID-19 misinformation using English fact-checked data via Google Translate API and a BERT-based sub-text aggregation model, without XAI. CrossFake-sub achieved 75% accuracy, outperforming monolingual and cross-lingual baselines. Translation errors, including rendering "COVID-19" as "new crown virus", are the dominant bottleneck. Padalko et al. [13] equally moved beyond binary classification by integrating BERT, BiLSTM, attention mechanisms, and BERTopic for narrative clustering, without XAI beyond topic model outputs. The system reached 99.47% accuracy and 0.9947 F1. Narratives are grouped into conspiracy theories, pseudoscientific claims, and political disinformation. The pipeline is text-only. Similarly,

Table 5. Summary of Reviewed Studies

Ref.	Dataset · language · size	Modalities	Model/method	Metrics	XAI	Best result	Resource
[2]	Facebook/Twitter · Luganda · not stated	Text	Weka classifiers (DMNBtext, LibSVM)	Acc, F1	None	Acc: 78.19%; F1: 76.9% (DMNBtext)	LRL
[3]	COVID-19-FNIR (translated) · Sepedi · not stated	Text	Zabantu (RoBERTa), SVM (+ augmentation)	Val. Acc	None	Acc: 99% (SVM, full augment.)	LRL
[7]	NELA-GT-2020 · English · not stated	Text	BERT + lexical/non-lexical metadata	Acc, F1, AUC	None	Acc: 82.4%; F1: 76.8%	HRL
[11]	ANTi-Vax · English · 15,000+	Text	BERT, XGBoost, LSTM	Acc, F1	None	Acc: 98%; F1: 98% (BERT)	HRL
[12]	ReCOVery, CoAID, FakeCovid · Chinese (via EN) · not stated	Text	BERT + CrossFake (Google Translate API)	Acc, Prec, Rec, F1	None	Acc: 75% (CrossFake-sub)	LRL
[41]	COVAXLIES · English · not stated	Text	TMD-GLP + COVID-Twitter-BERT-v2	Prec, Rec, F1	None	F1: 84.3% (TMD-GLP)	HRL
[43]	SA vs US disinformation · SA English · not stated	Text	LR, LSTM	Acc, F1	LIME	Acc: 90.6% (combined dataset)	HRL
[50]	CoAID translated · Sesotho · not stated	Text	SVM, LR, DT, GB, RF, NB (TF-IDF)	F1, Prec, Acc	None	Acc: 91% (SVM, Helsinki-NLP)	LRL
[51]	MarianMT COVID news · Xitsonga · not stated	Text	mBERT vs RF	Acc, F1, Prec, Rec	None	Acc: 98.7%; F1: 88% (mBERT)	LRL
[52]	Twitter hate speech · isiZulu · not stated	Text	SVM, LR, RF (TF-IDF)	Acc, Prec, Rec	None	Acc: 82.49%; Prec: 83.02% (SVM)	LRL

Kim et al. [61] tracked vaccine misinformation across Brazil, Indonesia, and Nigeria using XLM-RoBERTa with GPT-3-based augmentation for underrepresented themes, without XAI. XLM-R+_AUG achieved 97.8% macro F1 for vaccine relevance and 92.9% for misinformation relevance. Country-specific variation limits direct transfer between settings. In addition, Shang et al. [40] built CrowdAdapt to transfer COVID-19 knowledge to Mpox and Polio detection using a Graph-based Knowledge Encoder, Domain-invariant Representation Learning, and a Crowdsourcing-based Knowledge Updater. Expert crowd verification provided implicit validation. F1 gains over DETERRENT ranged from +1.2% to +9.1% for Mpox and +1.8% to +5.5% for Polio. Knowledge noise from crowd updates constrains adaptation quality.

B. Non-COVID-19 health misinformation

Studies related to health misinformation outside COVID-19 are discussed here. Aziz et al. [62] curated over 10,000 Arabic social media posts on diabetes, verified by medical experts through majority voting, and evaluated LSTM, DistilBERT, and GPT-4 without XAI beyond expert labelling. DistilBERT reached 96% accuracy; GPT-4 outputs aligned closely with expert judgements. Source biases in Arabic social media affect label reliability. Similarly, Khemani et al. [32] compared SVM, LR, RF, RNN, and LSTM for healthcare misinformation using TF-IDF, Bag of Words, and word embeddings, without XAI. LSTM reached 86% accuracy, ahead of RNN at 82% and all classical classifiers. Dataset quality constraints further accuracy gains in this domain. Marvin et al. [63] also applied LR to crowdsourced maternal and neonatal health content from Facebook, WhatsApp, Twitter, and LinkedIn, using ELi5 to clarify how the model identified medical sentiment. Training accuracy reached 93%; deployment precision reached 83%. General-purpose training data does not capture cultural specificity in health advice. On the same note, Zhao et al. [16] applied the Elaboration Likelihood Model to a Chinese autism forum, training ML classifiers on user behavioural features: discussion initiation frequency, social influence, and informational independence. Feature importance ranking served as interpretability; no formal XAI tool was used. Behavioural features outperformed text analysis; the model detected approximately 85% of medical misinformation. Platform-specific annotations constrain cross-platform transfer. Finally, Ramakrishnan and Balakrishnan [64] used lexical features such as word count, character length, and uppercase usage, with DT, RF, and SVM for bilingual health MID in English and Malay, without XAI. RF reached 95% accuracy and 92.8% F1 on the Malay dataset. Surface lexical features do not capture semantic or pragmatic deception.

D.3. Multimodal detection

This subsection focuses on studies dealing with hybrid modalities. Guerrero-Sosa et al. [24] combined DL multimodal feature extraction with fuzzy logic through the Granular Linguistic Model of Phenomena for TikTok disinformation, incorporating Big-5 personality trait analysis. Structured suspicion-level reports served as the explainability mechanism. Multimodal analysis reached 95% F1, outperforming unimodal methods. The framework assumes direct-to-camera speech, constraining applicability to produced or scripted video. Kim et al. [25] improved YouTube MID by incorporating metadata and thumbnails from platform-recommended related videos alongside Doc2Vec and CNN content representations, without XAI. YouRV reached 87.16% accuracy, improving on content-only baselines by 0.1–9.7%. Detection depends on YouTube recommendation logic, which can change.

Fu et al. [37] developed DVHAM using a dual-branch visual network, capturing global image structures and fine-grained local details, fused with hierarchical semantic text features through a gating mechanism, without XAI. DVHAM reached 91.5%, 92.4%, and 90.6% accuracy on Weibo, Twitter, and PHEME, respectively. Performance degrades

under adversarial image manipulation. Geethanjali and Valarmathi [10] built DeepCOVID-FakeSense pairing BERT-BiLSTM for text with GNN-CNN for images, without XAI. Text-only analysis reached 99.93% accuracy; multimodal fusion on the multilingual MM-COVID dataset reached 82.93%. Computational cost constraints deployment in lower-resource settings.

D.4. Model evolution

A. Classical ML and deep learning

Mohsen et al. [65] combined TF-IDF with the Empath tool to capture linguistic, psychological, and grammatical features, applied incremental learning for dataset adaptation, and released an open-source web framework. No XAI was used. LR outperformed DL baselines by approximately 4%, reaching 99% accuracy on Kaggle 2. Coverage is English only.

For DL, Jain et al. [31] applied HHO to select 80 discriminative linguistic features from a larger pool, then processed these through a CNN-BiLSTM hybrid without XAI. The system reached 98.89% on ISOT, 98.25% on Kaggle, and 98.56% on ConFake. Scalability to larger, noisier datasets is untested. Similarly, Ayyasamy et al. [26] integrated LSTM with a Convolutional Gaussian Perceptron Neural Network, allowing the model to represent uncertainty in deceptive text probabilistically, optimised by Moth-Flame Whale Optimisation, without XAI. The system reached up to 98% accuracy across four datasets, outperforming deterministic models by 3–8%. Training requires over 35 hours on high-end GPUs. Hashmi et al. [66] equally trained a bilingual CNN-LSTM with FastText embeddings for misogynistic content detection in English and Italian, applying LIME to visualise influential words. Incremental learning and fine-tuned FLAN-T5 supplemented the pipeline. CNN-LSTM reached 95% accuracy, ahead of ELECTRA_large at 88%. Training data is drawn from homogeneous platform sources, limiting cross-platform exposure.

B. Transformer-based and hybrid frameworks

Studies in this category are discussed as follows. Raza [7] combined BERT with lexical and non-lexical categorical and numeric metadata from NELA-GT-2020 using multi-head attention gating, without XAI. Faker reached 82.4% accuracy, 76.8% F1, and 80.4% AUC. This exceeded TriFN, Grover, GPT-2, and VGCN-BERT by 20–30%. Distant supervision labels introduce noise that the evaluation does not control for. Polu [9] combined BERT, RoBERTa, and XLNet with Graph Neural Networks and Graph Convolutional Networks, applying adversarial training via GAN-generated samples, back-translation, and synonym replacement, and used both LIME and SHAP for post-hoc explanation. RoBERTa reached 96.2% accuracy. Post-hoc explanations cannot fully capture the opacity of deep graph architectures.

Table 6. Summary of Reviewed Studies

Ref.	Dataset · language · size	Modalities	Model/method	Metrics	XAI	Best result	Resource
[13]	COVID19-FNIR · English · not stated	Text	BERT + BiLSTM + BERTopic	Acc, Prec, Rec, F1	None	Acc: 99.47%; F1: 0.9947	HRL
[25]	FVC, FYKE (YouTube) · Multi-lingual · not stated	Text, Image (thumb)	YouRV (Doc2Vec + CNN + metadata)	Acc, Prec, Rec, F1, AUC	None	Acc: 87.16%; +0.1–9.7% vs base	HRL
[26]	ISOT, Fakeddit, BuzzFeed · English · benchmark	Text	LSTM + CGPNN + MFWO	Acc, Prec, Rec, F1	None	Acc: 98% (BuzzFeed)	HRL
[31]	ISOT, Kaggle, ConFake, McIntire · English · benchmark	Text	HHO + CNN-BiLSTM	Acc, Prec, Rec, F1	None	Acc: 98.89% (ISOT)	HRL
[32]	Monant Medical · English · not stated	Text	SVM, LR, RF, RNN, LSTM	Acc, Prec, Rec, F1	None	Acc: 86% (LSTM)	HRL
[33]	Constraint@AAAI2021 · English · not stated	Text	CT-BERT + BiGRU (domain fine-tuned)	Acc, Prec, Rec, F1	None	F1: 98% (CT-BERT+BiGRU)	HRL
[44]	Facebook Amharic news · Amharic · 23,856	Text, Image	CNN + Attention-BiLSTM (late fusion)	Acc, Prec, Rec, F1	None	Acc: 98%	LRL
[52]	Weibo, Twitter, PHEME · English + Chinese · benchmark	Text, Image	DVHAM (dual-branch CNN + Transformer)	Acc, Prec, Rec, F1	None	Acc: 92.4% (Twitter)	HRL
[60]	COVID-19 English · English · not stated	Text	CNN-BiLSTM-AM + multi-head attention	Acc, F1, Rec	None	Acc: 93.6%	HRL
[65]	FA-KES, WELFake, Kaggle · English · benchmark	Text	TF-IDF + Empath + LR (incremental learning)	Acc, Prec, Rec, F1	None	Acc: 99% (Kaggle 2)	HRL

**LRL = low-resource language; HRL = high-resource language.

C. Optimisation, probabilistic modelling, and social context integration

Metaheuristic optimization and probabilistic modelling address two related problems, such as feature redundancy and uncertainty in deceptive text, that purely architectural choices do not resolve. Jain et al. [31], discussed in Section 4.5.1, demonstrated that systematic feature selection via HHO reduces noise before classification rather than relying on the model to learn around it. Ayyasamy et al. [26] show that Gaussian perceptron networks handle distributional uncertainty in misinformation text that deterministic architectures treat as a fixed signal.

Social and behavioural signals extend detection beyond what content analysis alone can achieve. Zhao et al. [16] find that user behavioural features, discussion initiation frequency, and social influence outperform text analysis in a health misinformation forum context, as discussed in Section 4.3.2. Akhtar et al. [38] show that bot-generated COVID-19 tweets carry measurably higher engagement rates than human-generated content, with approximately 3% versus under 1%, providing a behavioural signal separable from linguistic content. Weinzierl and Harabagiu [41] connect tweets through shared misinformation targets rather than user networks, representing a graph-based form of

social context modelling that sits between content analysis and propagation tracking, as discussed also in Section 4.3.1. In all, these studies indicate that optimisation, probabilistic design, and social feature integration each address failure modes that transformer fine-tuning alone does not cover.

E. Discussion

This section synthesizes findings from the 51 reviewed studies across MID in LRLs and the health domain. It covers eight thematic areas: resource creation; transformer dominance and its limits; multimodal detection; hybrid and metadata-enriched frameworks; task boundaries; explainability; contradictions and research gaps; and future directions.

E.1. Resource creation

Model performance across the 51 reviewed studies is bounded by dataset quality more than by architectural choice. Studies addressing Amharic, Bangla, Luganda, Sepedi, Sesotho, Xitsonga, isiZulu, Kurdish Sorani, Algerian Arabic, Dravidian, and Indic languages [1-3], [5], [8], [14], [15], [34], [39], [44], [48-52], [56] all required new corpora before any modelling could proceed. Where datasets were translated from English, augmentation strategies such as back-translation, synonym replacement, and random insertion partially compensated for the absence of native content. Translation quality is an unresolved constraint. Mokoena et al. [50] found Helsinki-NLP's Opus-MT more reliable than Google Translate for Sesotho, and CrossFake [12] traced classification noise directly to translation errors, including the rendering of COVID-19 as the new crown virus. Fig. 2 captures the distribution of the covered languages and domains, while Fig. 3 presents the distribution of studies across LRL and HRL and dataset size distribution (in log scale) by resource level.

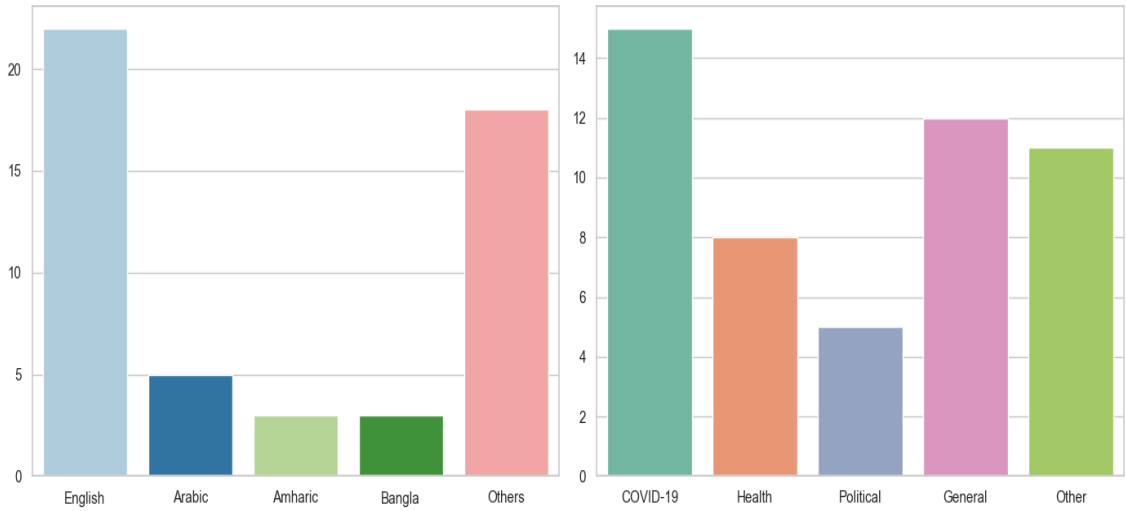

Figure 2. (a) Language coverage (b) Domain distribution.

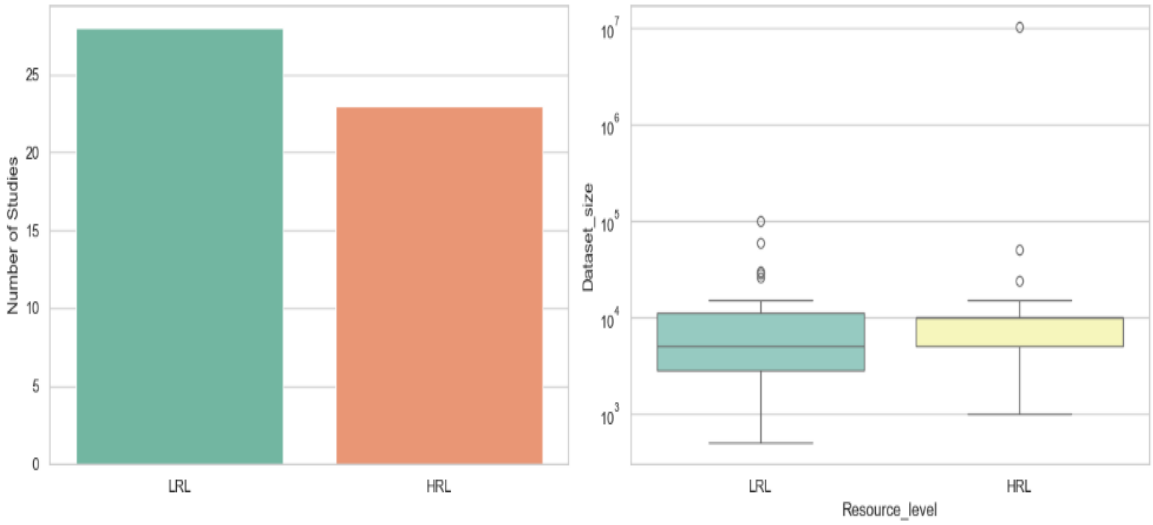

Figure 3. (a) Model types across studies (b) Performance range (min, avg., max) by model category.

However, scale compounds the problem. BanFakeNews-2.0 contains 60,000 items [8]; ArabFake covers only 2,500 [55]; the Tshivenda dataset is smaller still [49]. Performance figures from small corpora are not comparable to results on established English benchmarks. The resource creation is itself the contribution in these cases, but that distinction needs to be kept visible when interpreting reported metrics. Also, class imbalance cuts across LRL and general datasets alike. Bahurmuz et al. [54] address it through oversampling for Arabic rumour detection; Manaswe et al. [3] apply augmentation for Sepedi; Ayyasamy et al. [26] use probabilistic modelling to handle distributional uncertainty. No single approach dominates, and the reviewed studies do not compare these strategies directly. This is a methodological gap.

As shown in Fig. 2, English dominates the dataset; Arabic and Amharic account for a small number of additional studies, while most other languages appear in only one or two. Domain coverage concentrates on COVID-19 and general misinformation, with fewer studies addressing political or behavioural health misinformation. These gaps in language representation and domain diversity mean current research does not cover the full range of misinformation contexts. Fig. 3, slightly more studies address LRLs (28) than HRLs (23), yet dataset size distributions reveal a pronounced scale imbalance. LRL datasets are smaller and more variable; HRL datasets include several large-scale corpora. This pattern supports that performance constraints in LRL settings stem primarily from data availability and quality rather than model architecture, and results on small LRL datasets warrant cautious interpretation.

E.2. Transformer dominance and limits

Transformer-based models outperform classical ML and standard DL across nearly every language and domain in the reviewed set. XLM-RoBERTa, mBERT, AraBERT, MARBERT, and CT-BERT appear repeatedly as the strongest performers, across languages from Bengali [24] to Xitsonga [51] and Chinese [12], and across domains from vaccine misinformation [11], [33] to political fake news [7]. Domain-specific pre-training adds a measurable advantage: CT-BERT trained on COVID-Twitter data reached 98% F1 [33], outperforming vanilla BERT on the same task. However, three limits constrain this dominance. Computational cost is one. For instance, Ayyasamy et al. [26] reported training times exceeding 35 hours on high-end GPUs, a barrier that excludes deployment in exactly the low-resource institutional settings where LRL detection is most needed. Benchmark

accuracy is another: several studies report above 99% on balanced datasets such as ISOT and BuzzFeed [13], [26], [31], figures unlikely to hold on the noisy, imbalanced streams of live social media. Cross-lingual transfer is the third. CrossFake [13] achieves 75% accuracy on Chinese news via English translation, useful but degraded by translation errors and not generalisable to languages where machine translation quality is lower.

Three studies in the reviewed set use large language models (LLMs) rather than task-specific fine-tuned transformers. Shibu et al. [8] found BLOOM 560M reaches 89% macro F1 on Bangla; Kim et al. [61] use GPT-3 for data augmentation rather than classification; Aziz et al. [62] find GPT-4 outputs closely aligned with medical expert judgements on Arabic diabetes content. These results suggest LLMs are competitive but not superior to domain-fine-tuned transformers for classification tasks, a comparison that the reviewed studies do not make directly. As LLMs become more accessible, this comparison will require controlled investigation. Classical classifiers are not obsolete. SVM and RF remain competitive where data volume is limited, feature engineering is mature, or interpretability is required [3], [38], [64], [65]. The choice between classical and neural approaches should be governed by dataset size and deployment setting. Fig. 4 presents the distribution of model types and performance range across studies.

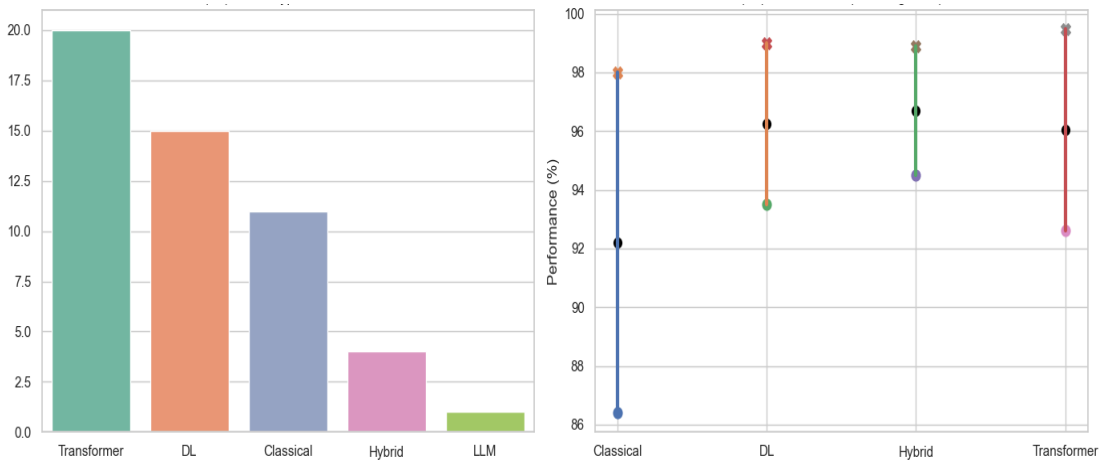

Figure 4. (a) Model types across studies **(b)** Performance range (min, avg., max) by model category.

In Fig. 4, transformer-based models are the most frequently used, followed by DL and classical ML approaches. Performance ranges show that hybrid and DL models achieve results comparable to transformers, and classical methods reach high performance in certain settings. As discussed above, transformer dominance in adoption does not translate to a consistent performance advantage; model selection remains dependent on data scale and deployment constraints. In the same vein, accuracy and FI-score dominated the metrics used to evaluate models'performance. The distribution is shown in Fig. 5.

E.3. Hybrid, metadata-enriched frameworks and multimodal detection

Several reviewed studies improve on text-only transformer baselines by incorporating non-textual signals. Raza [7] combines BERT with lexical and non-lexical metadata, categorical and numeric attributes from the NELA-GT-2020 dataset, exceeding TriFN, Grover, and VGCN-BERT by 20-30%. YouRV [25] incorporates YouTube recommendation metadata and video thumbnails, improving accuracy by up to 9.7% over content-only models. Sultana and Nishino [14] combine BERT with XGBoost ensemble methods for Bangla, and Mohsen et al. [65] pair TF-IDF with the Empath psychological

feature tool, outperforming DL baselines by approximately 4%. The pattern across these studies is consistent: metadata and auxiliary features add signal that text-only models miss, particularly for platform-specific propagation behaviour. Metadata integration is underexplored relative to its demonstrated contribution. Shehata et al. [55] extend this further through multitask learning. ArabFake simultaneously performs fake news detection, misinformation categorisation, and risk assessment within a single MARBERTv2 framework. Multitask architecture reduces annotation overhead and can produce richer outputs than binary classifiers, but no other reviewed study employs this approach, and its generalisability across languages and domains is not yet established.

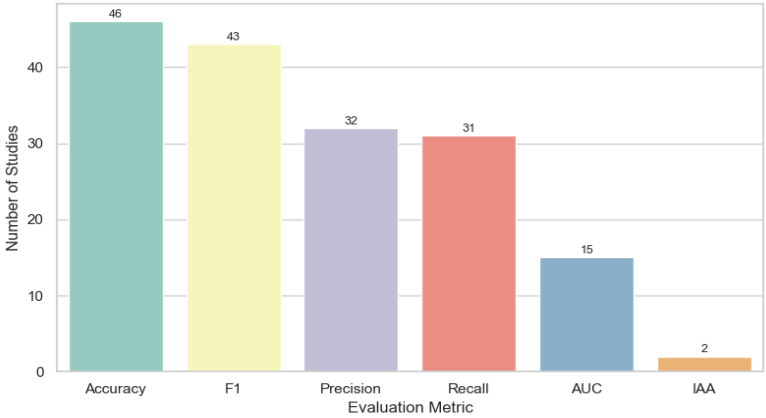

Figure 5. Distribution of evaluation metrics across studies

Multimodal fusion improves detection across all reviewed studies that employ it [10], [36], [37], [24], [25], [44]. Text remains the dominant modality: Bansal et al. [36] show through ablation that removing text from a multimodal Indic system reduces F1 by 9.3% for fake samples, while removing images produces a smaller drop. Images, audio, and video add a discriminative signal where misinformation exploits caption-image inconsistencies or platform-specific propagation patterns not visible in text alone. One study addresses audio directly. Imbwaga et al. [58] extract speech from YouTube videos using FFmpeg and Whisper AI transcription, then classify hate speech content in English and Kiswahili using RF, XGBoost, Bi-LSTM, BERT, and GPT-J-6B. RF achieved 98% accuracy on English and 94% on Kiswahili. This is the only reviewed study to implement audio-to-text pipeline detection, and it addresses hate speech rather than fake news, a task boundary distinction that matters. The pipeline is replicable for MID but has not been applied in that context. Fig. 6 captures the distribution of input modalities across studies.

Two constraints limit multimodal work more broadly. Multimodal corpora are small relative to platform-scale misinformation, and no reviewed study addresses audio-only or speech-based misinformation in any LRL setting. Adversarial vulnerability is the second constraint. Fu et al. [37] show DVHAM degrades under adversarial image manipulation; Polu [9] uses GAN-generated samples as a partial countermeasure. Neither resolves the problem. A system achieving 92% accuracy on clean data but failing on manipulated inputs is not deployment-ready against motivated adversaries.

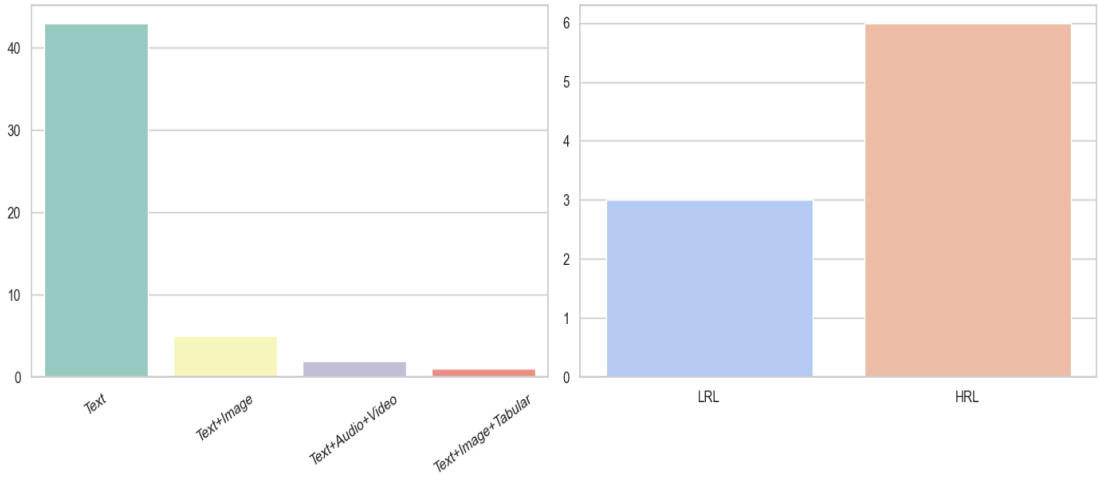

Figure 6. (a) Modality distribution **(b)** XAI adoption across resource levels.

As shown in Fig. 6, text-only approaches dominate the literature; few studies incorporate multimodal inputs such as images, audio, or video. XAI adoption remains limited overall and is least prevalent in LRL contexts (3 studies) compared to HRL settings (6 studies). As discussed in sections 5.3 and 5.4, multimodal approaches remain underexplored due to resource constraints, and explainability is absent where it is most needed.

E.4. Task boundaries and XAI

Task boundaries are discussed in terms of sarcasm, hate speech, and misinformation. Two reviewed studies addressed tasks adjacent to but distinct from fake news detection. Khan et al. [57] built the Urdu Sarcastic Tweets dataset and found that SVM achieves 84% F1, outperforming lexicon-based methods by a wide margin. Imbwaga et al. [58] addressed hate speech in English and Kiswahili. Both are included because sarcasm and hate speech share detection infrastructure: annotated social media data, text classification pipelines, and, in some cases, the same models, and because they co-occur with misinformation. This conflation is common in the field and is not always made explicit. Reviews that treat sarcasm, hate speech, and fake news as equivalent tasks risk overstating model generalisability, since each has distinct linguistic markers and requires different annotation protocols.

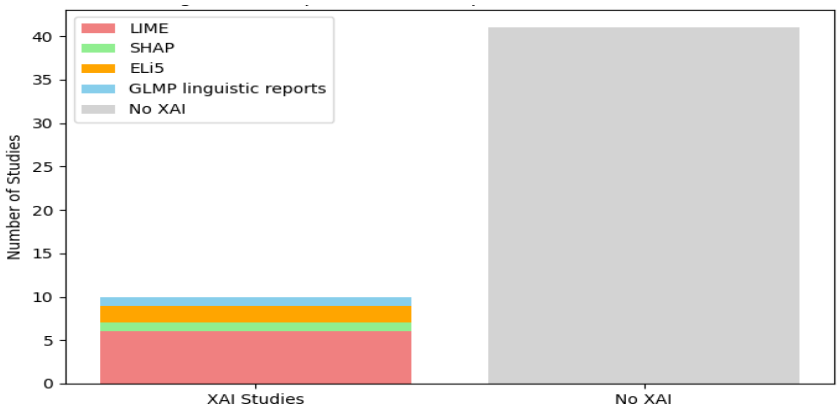

Figure 7. XAI adoption across studies

As shown in Fig. 7, across the studies reviewed, XAI adoption is real but narrow. LIME is the dominant method [1], [4], [42], [43], [58], [66]; SHAP appears in one study [9]; ELi5 in two [55], [63]. In every case, XAI is applied post-hoc to identify which words or features drove a specific prediction. This supports debugging and communicates individual decisions to non-expert users. It does not provide global explanations of model behaviour. No reviewed study formally tests whether explanations change how fact-checkers, healthcare providers, or policymakers make decisions. However, three specific gaps were observed. LIME's local explanations may not reflect the classifier's overall logic [58]. No reviewed study incorporates adaptive or temporal XAI. Several studies in low-resource settings, including the Tshivenda work [49], use no XAI at all, precisely the contexts where opaque automated decisions are most likely to go unchallenged. Social and behavioural signals offer a complementary form of interpretability. Zhao et al. [16] find that user behavioural features, such as discussion initiation frequency, outperform linguistic analysis for health MID in a Chinese autism forum. Akhtar et al. [38] show that bot-generated misinformation tweets have approximately 3% higher engagement than human-generated tweets. Wet and Marivate [43] use LIME to expose stylistic cues: repetitive phrasing and exaggerated sentiment, in South African English disinformation. In all, these findings suggest that social context, behavioural patterns, and stylistic signals carry information that text-only XAI methods do not surface. Their integration into explainability frameworks is underdeveloped.

E.5. Contradictions

Three patterns of contradictory findings demand attention. On retrieval method: Weinzierl and Harabagiu [41] found that classical BM25 scoring captured a broader range of vaccine misinformation tweets than BERT-based similarity. This is because keyword matching outperforms semantic matching for specific myths containing rare terminology. This contradicts the general finding of transformer superiority and suggests that dataset construction methodology inflates reported results. A model trained on BERT-retrieved data will appear stronger when evaluated on BERT-retrieved test data.

On augmentation: Manaswe et al. [3] reported substantial accuracy gains from back-translation and synonym replacement for Sepedi but also reported that higher augmentation volumes do not consistently raise quality and can introduce noise that degrades performance. The relationship between augmentation strategy, volume, and downstream accuracy is not well controlled in the reviewed studies. Similarly, on lexical features, Ramakrishnan and Balakrishnan [64] found that real news contains more uppercase letters and stop words, attributable to formal publication style. Other studies treat excessive capitalisation as a marker of fake news. Both can be correct in different language and platform contexts, but they cannot be reconciled directly, and lexical classifiers trained in one setting should not be applied elsewhere without validation.

E.6. Research gaps and possible future directions

Language coverage remains the field's most visible structural gap. Over 67% of reviewed studies focus on English [22]. African languages beyond Amharic and the four South African languages covered here, including Hausa, Yoruba, Tigrinya, and Somali, have no dedicated MID resources. Arabic dialectal variation is addressed by only one study [5]. Community-driven annotation, where native speakers contribute directly to corpus construction, offers a more durable path than translation pipelines alone. Audio-based misinformation in LRL settings is unaddressed as well. Imbwaga et al. [58] demonstrated that audio-to-text pipelines are viable for hate speech detection in Kiswahili, but no reviewed study applies this approach to misinformation in any LRL. Spoken misinformation circulates through WhatsApp voice messages and similar channels in many LRL communities, a distribution channel that text-and-image systems cannot reach.

Cross-domain transfer fails more often than it succeeds. Wet and Marivate [43] show that a model trained on US English disinformation achieves poor accuracy on South African English content due to stylistic differences, and combining both datasets raises accuracy to 90.6%. This is a different problem from cross-lingual transfer and is underexamined: even within English, domain and stylistic gaps degrade performance substantially. All 51 reviewed systems are static. Misinformation tactics evolve: bot behaviour adapts to evade detection [38], health crises generate new terminology [40], and platform recommendation logic shifts [25]. Mohsen et al. [65] showed incremental learning as a partial response, but no reviewed study builds a system that adapts continuously without full retraining.

Deployment evidence is absent. No reviewed study provides evidence on performance inside a live social media moderation pipeline. Adversarial vulnerability is acknowledged by Polu [9] and Fu et al. [37] but not resolved by either. Systems claiming 98-99% benchmark accuracy should be evaluated against adversarially generated inputs before deployment claims are made. No shared evaluation protocol exists across the reviewed studies. Metrics, dataset splits, baseline comparisons, and validation strategies differ substantially, making cross-study comparisons unreliable. A unified benchmarking framework covering LRL, health-domain, and multimodal settings would allow the field to measure progress more accurately. Bias from politically affiliated fact-checking sources [19], privacy risks in user-level behavioural modelling [16],[38], and the risk of automated systems suppressing legitimate speech are acknowledged in limitations sections across several studies, but not investigated empirically. These concerns require direct investigation as detection systems move toward deployment.

F. Proposed LRLM-EF

In this study, we reviewed 51 studies that cannot be directly compared. They highly differ in language, domain, dataset size, modality, method family, and the metrics used to report results. While some studies build new corpora, others fine-tune on existing benchmarks. In addition, some report accuracy only; others report F1 by class, etc. In this case, a study achieving 99% accuracy on a translated COVID-19 corpus in Sepedi cannot be ranked against one achieving 71% F1 on a few-shot Algerian Arabic classifier without a shared scoring basis. Thus, the absence of standardised evaluation is not incidental, but it reflects a structural gap in the field that this review identifies as the most consequential obstacle to progress. To address this gap, this review proposes the LRL misinformation evaluation framework (LRLM-EF). This is a five-criterion scoring protocol for assessing and comparing MID studies across languages and domains. LRLM-EF is applied retrospectively to all 51 reviewed studies. Its criteria are derived directly from the four unresolved gaps identified in Section 4, and the modality distribution revealed in 4.1.

We used LRLM-EF to score each study on five criteria. The maximum total score is 11 points. Table 7 defines each criterion, its scoring range, and the rubric applied.

Table 7. LRLM-EF scoring criteria and rubric

ID	Criterion	Range	Scoring rubric
C1	Resource coverage	0-2	0 = HRL/English only · 1 = LRL with existing corpus · 2 = LRL with new corpus constructed
C2	Modality scope	0-2	0 = text only · 1 = text + image · 2 = text + audio or video
C3	Adversarial robustness	0-2	0 = none · 1 = cross-domain transfer tested · 2 = explicit adversarial evaluation reported
C4	Explainability	0-3	0 = none · 1 = post-hoc XAI only · 2 =

	depth		intrinsic XAI · 3 = user decision-impact evaluated
C5	Evaluation standardisation	0-2	0 = proprietary setup only · 1 = public dataset used · 2 = public dataset + standard metrics reported
Maximum total score		11	C1×2 + C2×2 + C3×2 + C4×3 + C5×2. Studies scoring ≥8 meet the threshold for deployment-ready evidence claims.

As shown in Table 7:

- C1 indicates the resource coverage scores for whether the study addresses a documented under-resourced language or domain. Dataset construction for a new LRL corpus receives the maximum score of 2, reflecting the field's most acute need. Studies that operated exclusively on English or high-resource benchmarks score 0.
- C2 is the modality scope scores whether the study extends beyond text. The review finds that 86% of studies use text alone. C2 penalises this monoculture without excluding text-only work: a study with strong text-only contributions still scores on C1, C3, C4, and C5.
- C3 represents the adversarial robustness scores, whether the model is evaluated against adversarial or out-of-distribution inputs before deployment claims are made. No reviewed study reports explicit adversarial evaluation. C3 scores 0 across the entire corpus, making it the most uniformly absent criterion in the field.
- C4 represents the explainability depth scores of the XAI application and its evaluation firmness. The three-point range distinguishes absence of XAI, post-hoc application (10 studies achieve), intrinsic methods (none achieve), and user decision-impact evaluation (none achieve). The higher weight of C4 reflects the field's particular deficit here: deployment in public health contexts requires explainability that users can act on, not only attention maps that researchers can inspect.
- C5 is the evaluation standardisation score, indicating whether results are reported in a way that allows cross-study comparison. A study using a public dataset and reporting standard metrics (precision, recall, F1 by class) scores a maximum of 2. Studies using proprietary or translated corpora with single-metric reporting score 0, regardless of the accuracy achieved.

In the course of the review, LRLM-EF was applied to all 51 reviewed studies. Table 8 reports scores for a representative subset. The score distribution reveals two findings. First, no study in the corpus scores above 6 out of 11. The median score is 4. This is not a judgment on study quality within each study's stated scope; several studies are technically strong. It reflects that the field has not yet converged on a shared evaluation standard, and that most studies optimise for one or two dimensions while leaving others unaddressed. Second, C3 scores 0 across all 51 studies. Adversarial evaluation before deployment claims is entirely absent from the reviewed literature. Given that several studies make explicit deployment recommendations, specifically in health misinformation settings, this is the field's most urgent unresolved gap.

Table 8. LRLM-EF scores for a representative subset of reviewed studies

Study	Dataset · language	C1	C2	C3	C4	C5	Total /11
[2] Gemeda et al.	Ethio-Fake · Amharic	2	0	0	1	1	4
[3] Shibu et al.	BanFakeNews-2.0 · Bangla	2	0	0	0	2	4

[4] Bansal et al.	MMIFND · Multi-Indic	2	2	0	0	2	6
[6] Polu	LIAR/FakeNewsNet · English	0	0	0	2	2	4
[14] Imbwaga et al.	YouTube · EN+Kiswahili	1	2	0	1	1	5
[18] Salh & Nabi	KDFND · Kurdish Sorani	2	0	0	0	1	3
[20] Borgohain et al.	Dravidian_Fake · 4 langs	2	0	0	0	2	4
[28] Guerrero-Sosa et al.	TikTok · EN+Spanish	0	2	0	1	1	4
[37] Manaswe et al.	COVID-19-FNIR · Sepedi	2	0	0	0	1	3
[47] Geto et al.	Facebook · Amharic	2	2	0	0	1	5
[53] Alghamdi et al.	Constraint@AAAI · English	0	0	0	0	2	2

Note: C3 scores 0 for all studies in the full corpus. The column is retained to make the gap explicit.

Each LRLM-EF criterion addresses one of the four gaps identified in Section 4, as follows:

- Gap 1: missing language and domain resources → C1: Resource coverage
- Gap 2: adversarial evaluation absent before deployment claims → C3: Adversarial robustness
- Gap 3: XAI not evaluated for user utility → C4: Explainability depth
- Gap 4: no shared evaluation protocol → C5: Evaluation standardisation

However, C2 - modality scope is not mapped to one of the four gaps because it was identified through the taxonomy rather than the gap analysis. We included it because the 86% text-only rate and the complete absence of audio-based LRL detection represent a structural constraint that a shared evaluation framework should make visible.

G. Conclusion

This paper presents an analysis of peer-reviewed studies on MID, covering model development, dataset construction, multimodal integration, and explainability across LRL and health-domain settings. Findings reveal several patterns that hold across the reviewed set. Transformer-based architectures outperform classical ML and DL in most settings, though classical classifiers remain competitive where data is limited or interpretability matters. Multimodal fusion improves detection in every study that tests it, though the gains are modest and text remains the dominant modality. XAI adoption is real but shallow, as LIME, SHAP, and ELi5 appear in a few studies, all applied post-hoc to individual predictions, and none formally evaluate whether explanations change user decisions. In addition, other health-domain research is the second largest theme. COVID-19 and vaccine misinformation dominate, with dedicated datasets and strong model performance. Other health conditions like diabetes, maternal and neonatal health, and autism appear in isolated studies. Conditions such as HIV, malaria, and reproductive health, where misinformation carries severe consequences in low-resource communities, are absent from the reviewed set entirely. The evidence also contains contradictions that a results summary would miss. Classical retrieval outperforms neural similarity for datasets covering rare misinformation terminology, which means neural-retrieval-constructed datasets may overstate model performance. Augmentation volume does not reliably improve accuracy and can introduce noise. Lexical markers attributed to fake news in one context reflect formal publication conventions in another. These inconsistencies affect how datasets should be built and how results should be read.

Three structural gaps run through the full set. More than 67% of studies focus on English. Audio and speech-based misinformation in low-resource settings is untouched.

No reviewed system has been tested in a live moderation pipeline. Models reporting 98-99% accuracy on balanced benchmarks have not been evaluated under adversarial conditions or on the imbalanced, noisy streams of live social media. Several task boundaries also remain blurred. Sarcasm detection, hate speech, and fake news detection share infrastructure across the reviewed set but have distinct annotation requirements, linguistic markers, and failure modes. Treating them as equivalent inflates reported generalization. Four specific actions would address these gaps. Community-driven dataset construction is needed for languages with no annotated resources and for health domains beyond COVID-19 in low-resource contexts. Such languages in the African context include Hausa, Yoruba, and Somali, and dialectal Arabic beyond the Algerian case addressed by one study, while the health domain MID should focus on HIV, malaria, reproductive health, etc. Also, adversarial evaluation should precede any deployment claim. XAI should be tested for whether explanations actually change decisions by fact-checkers, clinicians, or policymakers, not merely whether they can be generated. A shared evaluation protocol covering LRL, health-domain, and multimodal settings would make cross-study performance comparisons meaningful. Each of these is tractable. None requires a new generation of models.

H. Acknowledgment

This research was supported by the Unit for Data Science and Computing (UDSC) and the Department of Computer Science at the North-West University Mafikeng campus, South Africa.

I. References

[1] Gemeda Yigezu, M., Mersha, M. A., Bade, G. Y., Kalita, J., Kolesnikova, O., & Gelbukh, A. Ethio-Fake: Cutting-Edge Approaches to Combat Fake News in Under-Resourced Languages Using Explainable AI. *Procedia Computer Science*, *244*, 133-142, 2023. https://doi.org/10.1016/j.procs.2024.10.186.

[2] Nabende, P., Kabiito, D., Babirye, C., Tusiime, H., & Nakatumba-Nabende, J. Misinformation detection in Luganda-English code-mixed social media text. *ArXiv, 2021*. https://arxiv.org/abs/2104.00124

[3] M. Manaswe, S. Rananga and T. Sindane, "Detecting COVID-19 Misinformation in Sepedi Short Texts Using Data Augmentation Techniques," 2025 IST-Africa Conference (IST-Africa), Nairobi, Kenya, 2025, pp. 1-7, doi: 10.23919/IST-Africa67297.2025.11060465.

[4] G. Joshi et al., "Explainable Misinformation Detection Across Multiple Social Media Platforms," in *IEEE Access*, vol. 11, pp. 23634-23646, 2023, doi: 10.1109/ACCESS.2023.3251892.

[5] A. Abdedaiem, A. H. Dahou, and M. A. Cheragui, “Fake News Detection in Low Resource Languages using SetFit Framework”, *IA*, vol. 26, no. 72, pp. 178–201, Sep. 2023.

[6] S. S. Nandgaonkar, J. Shaikh, G. B. Bhore, R. V. Kadam, and S. S. Gadhave, "Multilingual Misinformation Detection: Deep Learning Approaches for News Authenticity Assessment," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-6, doi: 10.1109/ICCCNT61001.2024.10724048.

[7] Shaina Raza. 2021. Automatic Fake News Detection in Political Platforms - A Transformer-based Approach. In Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political

Events from Text (CASE 2021), pages 68-78, Online. Association for Computational Linguistics.

[8] Shibu, H. M., Datta, S., Miah, M. S., Sami, N., Chowdhury, M. S., & Islam, M. S. From Scarcity to Capability: Empowering Fake News Detection in Low-Resource Languages with LLMs. *ArXiv,* 2025. https://arxiv.org/abs/2501.09604.

[9] O. R. Polu. AI-Based Fake News Detection Using NLP. *International Journal Of Artificial Intelligence & Machine Learning* (IJAIML), Vol. 3, Issue 2, July-December 2024. pp.231-239. https://doi.org/10.34218/IJAIML_03_02_019.

[10] R. Geethanjali and A. Valarmathi, "Deep Learning Approach to COVID-19 Misinformation on Social Media with Multimodal and Multilingual Deep COVID-Fake Sense Model," 2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS), Gobichettipalayam, India, 2024, pp. 570-580, doi: 10.1109/ICUIS64676.2024.10867214.

[11] Hayawi, K., Shahriar, S., Serhani, M., Taleb, I., & Mathew, S. ANTi-Vax: A novel Twitter dataset for COVID-19 vaccine misinformation detection. *Public Health*, *203*, 23-30, 2022. https://doi.org/10.1016/j.puhe.2021.11.022.

[12] Du, J., Dou, Y., Xia, C., Cui, L., Ma, J., & Yu, P. S. Cross-lingual COVID-19 Fake News Detection. *ArXiv,* 2021. https://arxiv.org/abs/2110.06495.

[13] Padalko H, Chomko V, Yakovlev S, Chumachenko D. A Novel Comprehensive Framework for Detecting and Understanding Health-Related Misinformation. *Information*, 2025; 16(3):175. https://doi.org/10.3390/info16030175.

[14] R. Sultana and T. Nishino, "Identification of Fake News in Low-Resource Language: A Deep Learning Approach," 2023 15th International Congress on Advanced Applied Informatics Winter (IIAI-AAI-Winter), Bali, Indonesia, 2023, pp. 322-329, doi: 10.1109/IIAI-AAI-Winter61682.2023.00066.

[15] Salh, D. Abubakr, and Nabi, R. Mala. "Kurdish Fake News Detection Based on Machine Learning Approaches", *Passer Journal of Basic and Applied Sciences*, 5, 2, 2023, 262-271. doi: 10.24271/psr.2023.380132.1226.

[16] Zhao, Y., Da, J., & Yan, J. Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches. *Information Processing & Management*, *58*(1), 102390, 2021. https://doi.org/10.1016/j.ipm.2020.102390.

[17] M. Q. Alnabhan and P. Branco, "Fake News Detection Using Deep Learning: A Systematic Literature Review," in *IEEE Access*, vol. 12, pp. 114435-114459, 2024, doi: 10.1109/ACCESS.2024.3435497.

[18] F. Gulzar Hussain, M. Wasim, S. Hameed, A. Rehman, M. Nabeel Asim, and A. Dengel, "Fake News Detection Landscape: Datasets, Data Modalities, AI Approaches, Their Challenges, and Future Perspectives," in *IEEE Access*, vol. 13, pp. 54757-54778, 2025, doi: 10.1109/ACCESS.2025.3553909.

[19] O. Bashaddadh, N. Omar, M. Mohd, and M. Nor Akmal Khalid, "Machine Learning and Deep Learning Approaches for Fake News Detection: A Systematic Review of Techniques, Challenges, and Advancements," in *IEEE Access*, vol. 13, pp. 90433-90466, 2025, doi: 10.1109/ACCESS.2025.3572051.

[20] S. Pal and D. Kumar, "Comprehensive Literature Review on Fake News Detection: Techniques, Challenges and Emerging Trends," 2024 1st International Conference on Advances in Computing, Communication and

Networking (ICAC2N), Greater Noida, India, 2024, pp. 1277-1283, doi: 10.1109/ICAC2N63387.2024.10894974.

[21] A. Bodaghi, K. A. Schmitt, P. Watine, and B. C. M. Fung, "A Literature Review on Detecting, Verifying, and Mitigating Online Misinformation," in *IEEE Transactions on Computational Social Systems*, vol. 11, no. 4, pp. 5119-5145, Aug. 2024, doi: 10.1109/TCSS.2023.3289031.

[22] Rananga, S., Isong, B., Modupe, A., & Marivate, V. Misinformation Detection: A Review for High and Low-Resource Languages. *Journal of Information Systems and Informatics*, 6(4), 2892–2922, 2024.

[23] B. Alkomah and F. Sheldon, "Advancements in Fake News Detection Using Machine and Deep Learning Models: Comprehensive Literature Review," 2023 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2023, pp. 845-852, doi: 10.1109/CSCI62032.2023.00142.

[24] Guerrero-Sosa, J.D.T., Montoro-Montarroso, A., Romero, F.P. et al. A new hybrid intelligent approach for multimodal detection of suspected disinformation on TikTok. *Multimed Tools Appl* 85, 37, 2026. https://doi.org/10.1007/s11042-026-21304-8

[25] Kim, J., Shin, Y., Jung, G., & Ahn, H. A hybrid detection method for YouTube fake news using related video data. *Engineering Applications of Artificial Intelligence*, *156*, 111130, 2025. https://doi.org/10.1016/j.engappai.2025.111130.

[26] Ayyasamy RK, Ponnusamy C, Bhargavi KN, Cherukuvada S, Babu GC, Amutha S, Gamu DT. A hybrid deep learning framework for fake news detection using LSTM-CGPNN and metaheuristic optimization. *Sci Rep.* 2025 Nov 24;15(1):41522. doi: 10.1038/s41598-025-25311-x. PMID: 41285851; PMCID: PMC12644812.

[27] E. Tata, J. Ajdari, and N. Besimi, "Fake News Detection: A Comprehensive Survey," 2023 46th MIPRO ICT and Electronics Convention (MIPRO), Opatija, Croatia, 2023, pp. 309-314, doi: 10.23919/MIPRO57284.2023.10159859.

[28] R. Raj, C. Sharma, R. Uttara, and C. R. Animon, "A Literature Review on Clickbait Detection Techniques for Social Media," 2024 11th International Conference on Reliability, *Infocom Technologies and Optimization (Trends and Future Directions*) (ICRITO), Noida, India, 2024, pp. 1-5, doi: 10.1109/ICRITO61523.2024.10522359.

[29] Schlicht, I.B., Fernandez, E., Chulvi, B. *et al.* Automatic detection of health misinformation: a systematic review. *J Ambient Intell Human Comput* 15, 2009-2021, 2024. https://doi.org/10.1007/s12652-023-04619-4.

[30] Feng X, Luo J, Yang Y, El Baz D, Shi L. Health Misinformation Detection: Approaches, Challenges, and Opportunities. *Inquiry*. 2025 Jan-Dec;62:469580251384784. doi: 10.1177/00469580251384784. Epub 2025 Nov 4. PMID: 41189452; PMCID: PMC12589804.

[31] Jain, M.K., Gopalani, D., & Meena, Y.K. Hybrid CNN-BiLSTM model with HHO feature selection for enhanced fake news detection. *Soc. Netw. Anal. Min.* 15, 43, 2025. https://doi.org/10.1007/s13278-025-01455-6

[32] B. Khemani, S. Divekar, P. Jain, S. Gupta, M. Chouk, and S. Patil, "Unmasking Healthcare Misinformation RNN-Based Detection," 2024 IEEE International

Conference on Computing, Power and Communication Technologies (IC2PCT), Greater Noida, India, 2024, pp. 1534-1539, doi: 10.1109/IC2PCT60090.2024.10486261.

[33] Alghamdi J, Lin Y, Luo S. Towards COVID-19 fake news detection using transformer-based models. Knowl. *Based Syst.* 2023 Aug 15;274:110642. doi: 10.1016/j.knosys.2023.110642. Epub 2023 May 19. PMID: 37250528; PMCID: PMC10197436.

[34] Raja, E., Soni, B., & Borgohain, S. K. Fake news detection in Dravidian languages using transfer learning with adaptive finetuning. *Engineering Applications of Artificial Intelligence*, *126*, 106877, 2023. https://doi.org/10.1016/j.engappai.2023.106877.

[35] M. I. Elias, Y. Mahmud, S. Mutalib, S. N. Kamaliah Kamarudin, R. Maskat and S. A. Rahman, "Fake News Prediction Using Hybrid Model-Systematic Literature Review," 2023 4th International Conference on Artificial Intelligence and Data Sciences (AiDAS), IPOH, Malaysia, 2023, pp. 281-286, doi: 10.1109/AiDAS60501.2023.10284628.

[36] Bansal, S., Singh, N. S., Dar, S. S., & Kumar, N. MMCFND: Multimodal Multilingual Caption-aware Fake News Detection for Low-resource Indic Languages. *ArXiv*, 2024. https://arxiv.org/abs/2410.10407.

[37] Fu X, Zhang Z, Sun Y, et al. Dual-Branch Hybrid Visual Networks and Hierarchical Adaptive Fusion Strategy: An Effective Multimodal Fake News Detection Model. *The European Journal on Artificial Intelligence*. 2025;38(4):543-561. doi:10.1177/30504554251351227.

[38] M. M. Akhtar, I. Karunanayake, B. Sharma, R. Masood, M. Ikram, and S. S. Kanhere, "Towards Automatic Annotation and Detection of Fake News," 2023 IEEE 48th Conference on Local Computer Networks (LCN), Daytona Beach, FL, USA, 2023, pp. 1-9, doi: 10.1109/LCN58197.2023.10223359.

[39] Kar, D., Bhardwaj, M., Samanta, S., & Azad, A. P. (2020). No Rumours Please! A Multi-Indic-Lingual Approach for COVID Fake-Tweet Detection. *ArXiv*. https://arxiv.org/abs/2010.06906.

[40] Shang, L., Zhang, Y., Yue, Z., Choi, Y., Zeng, H., & Wang, D. A Domain Adaptive Graph Learning Framework to Early Detection of Emergent Healthcare Misinformation on Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, *18*(1), 1408-1421, 2024. https://doi.org/10.1609/icwsm.v18i1.31398.

[41] Weinzierl, M. A., & Harabagiu, S. M. Automatic detection of COVID-19 vaccine misinformation with graph link prediction. *Journal of Biomedical Informatics*, *124*, 103955, 2021. https://doi.org/10.1016/j.jbi.2021.103955.

[42] Ibrahim Aboulola, O., Umer, M. Novel approach for Arabic fake news classification using embedding from large language features with CNN-LSTM ensemble model and explainable AI. *Sci Rep* 14, 30463 (2024). https://doi.org/10.1038/s41598-024-82111-5.

[43] H. de Wet and V. Marivate, "Is it Fake? News Disinformation Detection on South African News Websites," 2021 IEEE AFRICON, Arusha, Tanzania, United Republic of, 2021, pp. 1-6, doi: 10.1109/AFRICON51333.2021.9570905.

[44] Geto, A.D., Emiru, E.D., Seid, N.E. *et al.* Multimodal based Amharic fake news detection using CNN and attention-based BiLSTM. *Sci Rep* 15, 34447 (2025). https://doi.org/10.1038/s41598-025-17579-w.

[45] B. R. P. Damoto et al., "Deep Learning and Ensemble Approaches to Misinformation Detection in Digital News: A Systematic Review," 2024 IEEE 10th Information Technology International Seminar (ITIS), Surabaya, Indonesia, 2024, pp. 151-156, doi: 10.1109/ITIS64716.2024.10845709.

[46] M. J. Page *et al.*, "PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews," *BMJ,* vol. 372, 2021.

[47] Sarkis-Onofre, R., Catalá-López, F., Aromataris, E. *et al.* How to properly use the PRISMA Statement. *Syst Rev* 10, 117 (2021). https://doi.org/10.1186/s13643-021-01671-z.

[48] Gereme F, Zhu W, Ayall T, Alemu D. Combating Fake News in "Low-Resource" Languages: Amharic Fake News Detection Accompanied by Resource Crafting. Information. 2021; 12(1):20. https://doi.org/10.3390/info12010020.

[49] M. Mukwevho, S. Rananga, M. S. Mbooi, B. Isong, and V. Marivate, "Building a Dataset for Misinformation Detection in the Low-Resource Language," 2024 IST-Africa Conference (IST-Africa), Dublin, Ireland, 2024, pp. 01-09, doi: 10.23919/IST-Africa63983.2024.10569562.

[50] L. Mokoena, S. Rananga, and M. Mbooi, "Misinformation Detection in COVID-19 News for Low-Resource Languages: A Sesotho Case Study," 2025 IST-Africa Conference (IST-Africa), Nairobi, Kenya, 2025, pp. 1-8, doi: 10.23919/IST-Africa67297.2025.11060466.

[51] M. G. Baloyi, S. Rananga and M. S. Mbooi, "Fine-Tuning a Machine Learning Model to Detect Covid-19 Misinformation in Xitsonga," 2025 IST-Africa Conference (IST-Africa), Nairobi, Kenya, 2025, pp. 1-10, doi: 10.23919/IST-Africa67297.2025.11060489.

[52] M. R. Sethosa, S. Rananga and M. S. Mbooi, "Hate Speech Detection in isiZulu: A Semantic Approach," 2025 IST-Africa Conference (IST-Africa), Nairobi, Kenya, 2025, pp. 1-9, doi: 10.23919/IST-Africa67297.2025.11060561.

[53] Hossain, M. R., Hoque, M. M., Siddique, N., & Dewan, M. A. A. (2024). AraCovTexFinder: Leveraging the transformer-based language model for Arabic COVID-19 text identification. *Engineering Applications of Artificial Intelligence*, *133*, 107987. https://doi.org/10.1016/j.engappai.2024.107987.

[54] N. O. Bahurmuz, G. A. Amoudi, F. A. Baothman, A. T., Jamal, H. S. Alghamdi, and A. M. Alhothali, "Arabic Rumor Detection Using Contextual Deep Bidirectional Language Modeling," in *IEEE Access*, vol. 10, pp. 114907-114918, 2022, doi: 10.1109/ACCESS.2022.3217522.

[55] A. M. K. Shehata, M. Nasser Al-Suqri, N. Eldin Mohamed Elshaiekh Osman, F. Hamad, Y. Nasser Alhusaini, and A. Mahfouz, "ArabFake: A Multitask Deep Learning Framework for Arabic Fake News Detection, Categorization, and Risk Prediction," in *IEEE Access*, vol. 12, pp. 191345-191360, 2024, doi: 10.1109/ACCESS.2024.3518204.

[56] M. Sijanur Rahman, O. Sharif, A. Das, S. Afroze, and M. Moshiul Hoque, "FaND-X: Fake News Detection using Transformer-based Multilingual Masked

Language Model," 2022 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE), Naya Raipur, India, 2022, pp. 153-158, doi: 10.1109/WIECON-ECE57977.2022.10150461.

[57] Khan, S., Qasim, I., Khan, W., Khan, A., Khan, J. A., Qahmash, A., & Ghadi, Y. Y. An automated approach to identify sarcasm in low-resource languages. *PLOS ONE*, *19*(12), e0307186, 2024. https://doi.org/10.1371/journal.pone.0307186.

[58] Imbwaga, J.L., Chittaragi, N.B., & Koolagudi, S.G. Explainable hate speech detection using LIME. *Int J Speech Technol* 27, 793–815 (2024). https://doi.org/10.1007/s10772-024-10135-3.

[59] Akhter, M., Hossain, S.M.M., Nigar, R.S. *et al.* COVID-19 Fake News Detection using Deep Learning Model. *Ann. Data. Sci.* 11, 2167–2198 (2024). https://doi.org/10.1007/s40745-023-00507-y.

[60] Xia, H., Wang, Y., Zhang, J. Z., Zheng, L. J., Kamal, M. M., & Arya, V. COVID-19 fake news detection: A hybrid CNN-BiLSTM-AM model. *Technological Forecasting and Social Change*, *195*, 122746, 2023. https://doi.org/10.1016/j.techfore.2023.122746.

[61] Jongin Kim, Byeo Rhee Bak, Aditya Agrawal, Jiaxi Wu, Veronika Wirtz, Traci Hong, and Derry Wijaya. 2023. COVID-19 Vaccine Misinformation in Middle Income Countries. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3903–3915, Singapore. Association for Computational Linguistics.

[62] E. H. Aziz, A. A. S. Ibrahim, A. M. Said, M. Hossam, and M. T. Waheed, "Detecting Diabetes Misinformation: Leveraging Medical Expertise and NLP Models for Effective Detection," 2024 Intelligent Methods, Systems, and Applications (IMSA), Giza, Egypt, 2024, pp. 130-134, doi: 10.1109/IMSA61967.2024.10652616.

[63] G. Marvin, N. Hellen, J. Nakatumba-Nabende, and M. G. Rabiul Alam, "Trustworthy Medical Sentiment Detection for Maternal and Neonatal Healthcare," 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2023, pp. 575-580, doi: 10.1109/ICOEI56765.2023.10126020.

[64] K. Ramakrishnan and V. Balakrishnan, "Health Misinformation in the Covid-19 Era - Detecting Misinformation on Bi-lingual Corpora using Lexical Features," 2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), Maldives, Maldives, 2022, pp. 1-5, doi: 10.1109/ICECCME55909.2022.9988481.

[65] Mohsen F, Chaushi B, Abdelhaq H, Karastoyanova D, Wang K. Automated Detection of Misinformation: A Hybrid Approach for Fake News Detection. Future Internet. 2024; 16(10):352. https://doi.org/10.3390/fi16100352.

[66] Hashmi, E., Yayilgan, S.Y., Yamin, M.M. *et al.* Enhancing misogyny detection in bilingual texts using explainable AI and multilingual fine-tuned transformers. *Complex Intell. Syst.* 11, 39, 2025. https://doi.org/10.1007/s40747-024-01655-1.