International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

Zero-shot cross-lingual transfer for disinformation detection in low-resource Indic languages

Khadijatul Kubra Jessica * and Rishita Chakma

Department of Computer Science and Engineering, Rangamati Science and Technology University, Rangamati 4500, Bangladesh.

International Journal of Science and Research Archive, 2026, 18(03), 1391-1400

Publication history: Received on 17 February 2026; revised on 23 March 2026; accepted on 26 March 2026

Article DOI: https://doi.org/10.30574/ijsra.2026.18.3.0600

Abstract

Disinformation poses a severe threat in low-resource linguistic settings lacking annotated data and robust detection infrastructure. This study investigates the efficacy of multilingual transformers for zero-shot and few-shot cross-lingual disinformation detection, transferring knowledge from English to four Indic languages (Bangla, Hindi, Malayalam, and Tamil). Zero-shot evaluations reveal that large-scale models enable substantial knowledge transfer—with XLM-RoBERTa achieving a peak Macro-F1 of 0.726 on Bangla—while traditional TF-IDF baselines fail entirely, underscoring the necessity of deep contextual embeddings. Furthermore, few-shot adaptation utilizing merely 200 target-language samples yields profound improvements, elevating MuRIL's performance to an average Indic Macro-F1 of 0.805 and 0.945 on Hindi. Post-hoc explainability analyses utilizing LIME and SHAP confirm that these models successfully identify genuine, cross-lingual stylistic cues of deception (e.g., sensationalism and urgency markers) rather than relying on spurious memorization. Ultimately, this research demonstrates that combining multilingual transformers with minimal target supervision provides a highly effective, scalable framework for combating disinformation in resource-constrained environments.

Keywords: Cross-Lingual Transfer; Disinformation Detection; Fake News; Low-Resource Languages; Multilingual Transformers; Zero-Shot Learning

1. Introduction

The rapid proliferation of disinformation across digital platforms is a critical modern challenge, spreading exponentially faster than verified facts by exploiting psychological vulnerabilities and algorithmic amplification [1-4]. Democratized content creation further exacerbates this issue, allowing unverified narratives to reach millions globally [5, 6]. The ramifications are profound, ranging from democratic disruption to severe public health crises like the COVID-19 "infodemic" [7-10]. Additionally, the integration of multimodal elements and generative AI produces highly sophisticated synthetic media that easily bypass traditional verification mechanisms [11-13].

Automated disinformation detection has advanced significantly through deep learning and natural language processing (NLP), with transformer-based architectures (e.g., BERT, RoBERTa) establishing new text classification benchmarks [14, 15]. However, these advancements primarily benefit high-resource languages like English. Consequently, low-resource languages—such as the Indic family (Bangla, Hindi, Malayalam, Tamil)—remain disproportionately underrepresented [16, 17] and highly vulnerable to localized campaigns despite serving over a billion speakers [18, 19]. The primary barriers to robust detection in Indic languages are the severe scarcity of annotated datasets [20, 21] and unique linguistic complexities, including diverse scripts, rich morphology, and frequent code-mixing [19, 22].

* Corresponding author: Khadijatul Kubra Jessica

While multilingual transformers (e.g., mBERT, XLM-RoBERTa) and Indic-specific models (e.g., MuRIL) enable zero-shot and few-shot cross-lingual transfer learning [23-25], they often suffer performance degradation due to domain mismatches and language-specific stylistic nuances. Furthermore, the "black-box" nature of advanced transformers hinders transparency and accountability, which are paramount in high-stakes disinformation detection [26, 27]. The integration of Explainable Artificial Intelligence (XAI) techniques, such as LIME and SHAP, is therefore essential to demystify model predictions and identify key linguistic indicators.

To address these disparities, this study proposes a comprehensive cross-lingual disinformation detection framework tailored for low-resource Indic languages, focusing on Bangla. By leveraging English-trained multilingual transformers, this research systematically investigates transfer learning efficacy while integrating advanced XAI techniques to ensure the interpretability of cross-lingual deception detection.

The primary contributions of this paper are summarized as follows:

- **Comprehensive Benchmark Evaluation:** We present a systematic evaluation of zero-shot cross-lingual transfer efficacy for disinformation detection using advanced multilingual transformers from English to four low-resource Indic languages (Bangla, Hindi, Malayalam, and Tamil).
- **Ablation Study with Traditional Baselines:** We conduct a thorough comparative analysis against traditional machine learning baselines.
- **Few-Shot Adaptation Insights:** We investigate the impact of few-shot fine-tuning on the top-performing model, quantifying the performance gains achieved when utilizing highly limited target-language data.
- **Integration of Explainability (XAI):** We apply LIME and SHAP to interpret model predictions at both local (individual instance) and global (feature importance) levels, uncovering universal versus language-specific linguistic indicators of deception.
- **Qualitative Error Analysis:** We provide a comprehensive qualitative analysis of model predictions, highlighting the specific limitations of cross-lingual transfer, such as the handling of code-mixing and culturally nuanced narratives, offering actionable recommendations for future research in equitable AI.

2. Literature Review

2.1. Definitions and Societal Impact of Disinformation

Terms like "misinformation" (unintentional inaccuracies) [30], "disinformation" (deliberate deception) [28, 29], and "malinformation" (weaponized facts) [28] carry distinct implications. Due to the semantic dilution of "fake news" [7], researchers increasingly favor precise terminology [29, 30]. Disinformation severely impacts society by exacerbating polarization [31], influencing elections [7, 32], and hampering public health responses [8-10]. Furthermore, generative AI enables the large-scale production of multimodal synthetic media [11-13, 33]. These threats disproportionately affect low-resource linguistic contexts lacking adequate detection infrastructure [16, 17, 34].

2.2. Traditional Machine Learning in Deception Detection

Early automated detection relied on traditional classifiers (e.g., SVM, Random Forest) [35, 36, 37] utilizing manually engineered features across content [38, 39], user profiles [36, 40], and propagation dynamics [3, 41]. Deceptive content frequently exhibits distinct stylistic markers, including sensational language [38, 42] and specific psycholinguistic traits [38]. While these classifiers achieved 75–90% accuracy on early monolingual datasets like LIAR and FakeNewsNet [36, 39, 43], handcrafted features proved brittle, failing to generalize across diverse domains and languages [37].

2.3. Deep Learning and Transformer Architectures

Deep learning models, such as Bi-LSTMs [20, 44] and CNNs [45], improved feature extraction but struggled across languages due to their reliance on static word embeddings [46]. Subsequently, transformer architectures revolutionized detection via self-attention mechanisms [47, 48], with models like BERT and RoBERTa establishing new benchmarks [14, 15, 49].

2.3.1. Multilingual and Indic-Specific Models

To address cross-linguistic constraints, multilingual transformers like mBERT [14, 50] and XLM-RoBERTa [23] were developed. Despite their early success in cross-lingual transfer, generalized multilingual models frequently suffer from vocabulary fragmentation when applied to morphologically rich or low-resource languages [51].

2.3.2. Cross-Lingual Transfer Learning

Cross-lingual transfer learning has emerged as a vital paradigm for mitigating data scarcity, leveraging knowledge from resource-rich languages to facilitate robust target-language performance without extensive parallel corpora [22, 23, 52]. Zero-shot transfer achieves this by fine-tuning exclusively on source-language data before target-language evaluation.

2.3.3. Explainability in Deception Detection

Addressing the "black-box" nature of advanced transformers requires the integration of Explainable Artificial Intelligence (XAI) [53]. Techniques such as LIME and SHAP are vital for demystifying model decisions and uncovering linguistic indicators [54], though researchers emphasize they must be critically evaluated for vulnerability to adversarial perturbations [55].

3. Methodology

3.1. Overall Framework

This study employs a structured cross-lingual transfer learning framework designed to address the critical gap in disinformation detection for low-resource Indic languages. By leveraging deep contextual knowledge acquired from a high-resource language (English), the proposed methodology integrates zero-shot transfer, few-shot adaptation, systematic ablation, and post-hoc explainability to ensure robust, transparent, and equitable performance evaluation. Figure 1 illustrates the overall framework, with an emphasis on the flow from data preparation to evaluation and interpretability.

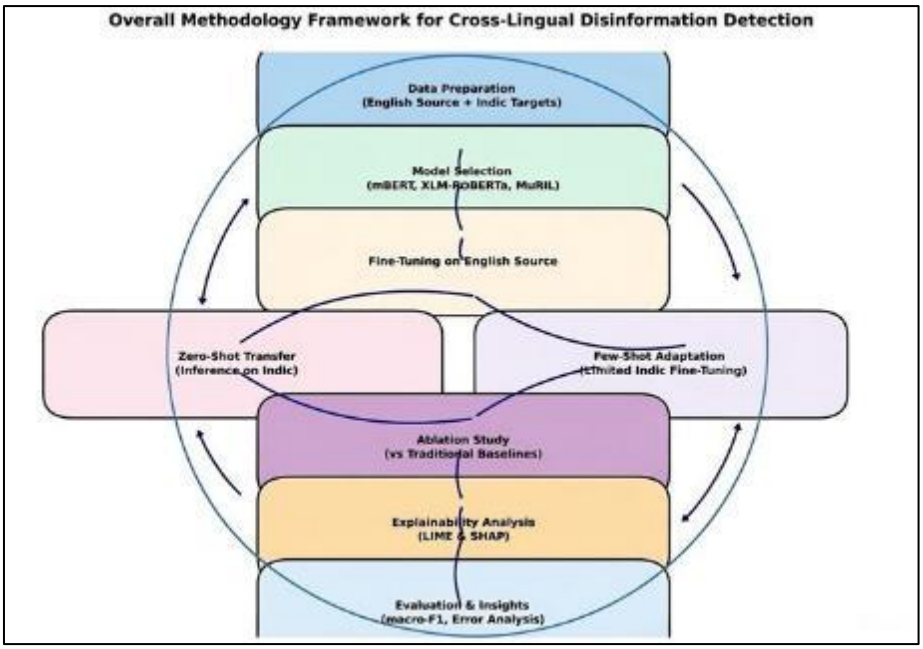

Figure 1 Overall Methodology Framework for Cross-Lingual Disinformation Detection in Indic Languages

The experimental pipeline consists of five primary stages:

- **Data Preparation:** Aggregation and standardized preprocessing of a large-scale English source dataset alongside four curated Indic target datasets (Bangla, Hindi, Malayalam, and Tamil).
- **Model Selection & Ablation:** Evaluation of advanced multilingual transformer models (mBERT, XLM-RoBERTa, and MuRIL) against traditional machine learning baselines (TF-IDF coupled with Logistic Regression, SVM, and Random Forest) to quantify the necessity of deep contextual representations.
- **Zero-Shot Transfer:** Fine-tuning the selected transformer models exclusively on the English dataset, followed by direct inference on the Indic test sets to evaluate baseline cross-lingual generalization.
- **Few-Shot Adaptation:** Strategic fine-tuning on highly constrained subsets of target-language data to simulate realistic, data-scarce annotation scenarios.

- **Explainability Analysis:** Application of LIME and SHAP to extract local and global interpretations of model predictions, identifying language-agnostic versus language-specific linguistic indicators of deception.

3.2. Datasets and Preprocessing

3.2.1. English Source Dataset

The English dataset serves as the high-resource foundation for cross-lingual transfer. To ensure adequate scale, linguistic diversity, and representation of contemporary disinformation strategies, the source corpus was constructed by aggregating three well-established repositories: LIAR, the ISOT Fake News Dataset, and FakeNewsNet. The raw data from these repositories underwent deduplication and standardization into a unified format comprising textual content (concatenated headlines and body text), binary labels (Real = 0, Fake = 1), and inferred domain metadata.

3.2.2. Indic Target Datasets

Target evaluation was conducted on four low-resource Indic languages. These datasets were strategically selected for their authenticity and representation of native disinformation patterns, such as culturally specific narratives and code-mixing:

- **Bangla:** Derived from the BanFakeNews 2.0 corpus, this subset comprises 25,800 balanced instances (12,900 real, 12,900 fake) covering politics, entertainment, and current events.
- **Hindi:** Sourced from the Hindi Fake and Real News corpus [59], the processed dataset contains 20,593 instances (10,300 real, 10,293 fake), heavily featuring the English code-mixing typical of South Asian digital media.
- **Malayalam:** Extracted from the DravidianLangTech-EACL2024 shared task, this dataset provides 5,091 instances (2,579 real, 2,512 fake) sourced directly from regional news domains.
- **Tamil:** Based on the Tamil Fake News Headlines corpus, this subset yields 5,148 instances (2,324 real, 2,824 fake) encompassing politics, sports, and technology.

All target datasets were preprocessed identically to the English source data—including text cleaning, normalization, and stratified splitting—to ensure strict cross-lingual consistency. The varying sizes and marginal imbalances of these datasets effectively mirror real-world, low-resource operational constraints.

3.3. Experimental Setup

3.3.1. Zero-Shot Cross-Lingual Transfer

The zero-shot protocol forms the core evaluation paradigm, assessing the models' capacity to generalize deception detection across linguistic boundaries without explicit target-language supervision. The multilingual transformers were fine-tuned exclusively on the English source dataset using a binary cross-entropy loss function. During inference, the fine-tuned models were applied directly to the Bangla, Hindi, Malayalam, and Tamil test sets. Input text was tokenized using the models' shared multilingual vocabularies, relying entirely on cross-lingual alignment within the embedding space to execute classification.

3.3.2. Few-Shot Adaptation

To quantify the performance gains achievable with minimal target-language supervision, few-shot adaptation experiments were conducted by extending the fine-tuning process from the zero-shot checkpoints. Two complementary configurations were designed to simulate practical annotation limitations:

- **Per-Language Adaptation:** A balanced subset of 200 labeled examples was sampled from each of the four Indic languages (800 total training instances). This simulates a scenario where minimal, uniformly distributed crowd-sourced annotations are available.
- **Bangla-Focused Adaptation:** Fine-tuning was restricted to 200 labeled examples exclusively from the Bangla dataset. This configuration investigates whether supervised adaptation in a single Indic language yields downstream performance improvements in other Indic languages via shared multilingual representations.

3.3.3. Implementation Details and Hyperparameters

To maintain the integrity of the zero-shot evaluation and prevent data leakage, hyperparameter tuning was conducted exclusively on the English validation set. A systematic grid search was utilized to identify optimal configurations. The learning rate was evaluated across:$\{1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}\}$

with batch sizes of $\{16, 32\}$. Models were trained for a maximum of 10 epochs, incorporating an early stopping mechanism governed by validation Macro-F1 with a patience of 3 epochs. Additional parameters included a 10% warmup proportion, a weight decay of 0.01, and default model dropout rates.

3.4. Evaluation Metrics

Model performance was quantitatively assessed utilizing standard binary classification metrics. Given the marginal class imbalances present in the target datasets, Macro-F1 was prioritized as the primary evaluation metric to ensure equitable representation of both the majority and minority classes. Additional reported metrics include Accuracy, per-class Precision, per-class Recall, standard F1-Score, and Weighted F1-Score.

3.5. Explainability (XAI) Framework

To mitigate the inherent opacity of advanced transformer architectures, post-hoc explainability techniques were integrated into the evaluation pipeline. Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) were selected for their complementary strengths in demystifying model decisions.

- **LIME Implementation:** LIME was utilized to generate instance-specific explanations by fitting a sparse linear model to locally approximate the transformer's behavior. For each instance, 5,000 perturbations were generated with 25% of tokens randomly masked, utilizing unigram and bigram bag-of-words features. The framework outputs the top 10 positive (deceptive) and negative (authentic) features.
- **SHAP Implementation:** Providing consistent feature attributions grounded in cooperative game theory, Kernel SHAP was applied for global summary plots (utilizing 1,000 background samples from the English training set), while Gradient SHAP was leveraged for localized, token-level explanations. A background dataset of 100 random English samples was utilized to establish expected values.

4. Results and Discussion

4.1. Quantitative Results

4.1.1. Zero-Shot Cross-Lingual Transfer

The zero-shot cross-lingual transfer experiments evaluated the capacity of models trained exclusively on English data to detect disinformation in Bangla, Hindi, Malayalam, and Tamil. As presented in Table 1, while all transformer models achieved strong baseline performance on the English source dataset (Macro-F1 approx. 0.90), their cross-lingual generalization varied significantly.

Table 1 Zero-Shot Macro-F1 Scores Across Models and Languages

Model	English	Bangla	Hindi	Malayalam	Tamil
mBERT	0.895	0.334	0.390	0.475	0.615
XLM-RoBERTa (large)	0.900	0.726	0.649	0.345	0.489
MuRIL	0.901	0.652	0.661	0.344	0.476

XLM-RoBERTa (large) demonstrated the most robust overall transfer to Indic languages, achieving the highest Macro-F1 on Bangla (0.726) and the highest average Indic score (0.552). This superior generalization is attributable to its massive pre-training scale and exposure to diverse, multilingual web corpora. MuRIL exhibited highly competitive and balanced results, notably leading in Hindi (0.661) and performing robustly on Bangla (0.652), reflecting the explicit benefits of its Indic-specific pre-training and transliteration augmentation. Conversely, mBERT displayed the weakest cross-lingual transfer, particularly on Bangla (0.334) and Hindi (0.390), aligning with its limited exposure to low-resource Indic data during its foundational pre-training.

4.1.2. Ablation Study: Transformers vs. Traditional Baselines

To quantify the necessity of deep contextual embeddings, an ablation study compared the transformer models against traditional machine learning baselines (Logistic Regression, Linear SVM, and Random Forest utilizing TF-IDF feature extraction).

Table 2 Ablation Study (Zero-Shot Macro-F1 Scores)

Model	English	Bangla
Logistic Regression (TF-IDF)	0.890	0.398
Linear SVM (TF-IDF)	0.880	0.388
Random Forest (TF-IDF)	0.889	0.334
mBERT	0.895	0.334
XLM-RoBERTa (large)	0.900	0.726
MuRIL	0.901	0.652

The ablation results (Table 2) reveal that while traditional baselines perform adequately on the English source language (Macro-F1 approx. 0.89), they fail entirely in a cross-lingual setting. On Bangla, the baselines drop to an average Macro-F1 of approximately 0.37, indicating virtually no effective knowledge transfer. This stark contrast underscores that surface-level lexical characteristics (n-grams) cannot bridge semantic alignments across disparate scripts and typologies. Multilingual transformers significantly outperform these baselines, confirming the critical role of contextual embeddings in enabling cross-lingual generalization.

4.1.3. Few-Shot Adaptation

To assess the viability of minimal-supervision scenarios, few-shot adaptation was evaluated under two configurations: a *per-language* setting (200 target-language samples per Indic language utilizing MuRIL) and a *Bangla-focused* setting (200 Bangla samples only, utilizing XLM-RoBERTa, to test positive transfer to related languages).

Table 3 Few-Shot Adaptation Results (Macro-F1 Scores)

Model	Setting	English	Bangla	Hindi	Malayalam	Tamil
MuRIL	Zero-Shot	0.901	0.652	0.661	0.344	0.476
MuRIL	Few-Shot (200/lang)	0.902	0.726	0.945	0.615	0.934
XLM-RoBERTa (large)	Zero-Shot	0.900	0.726	0.649	0.345	0.489
XLM-RoBERTa (large)	Few-Shot (200 Bangla only)	0.901	0.780	0.795	0.339	0.410

The per-language adaptation yielded substantial improvements across all Indic languages. Most notably, Hindi and Tamil experienced gains of +28.4% and +45.7% respectively, demonstrating that combining Indic-specific pre-training with highly limited target supervision effectively bridges the cross-lingual gap. The Bangla-focused adaptation provided moderate downstream gains for closely related languages (Bangla +5.4%, Hindi +14.6%) but resulted in negligible or negative transfer for morphologically distant Dravidian languages (Malayalam -0.6%, Tamil -7.9%). Importantly, source-language (English) performance remained stable across all adaptations, indicating no catastrophic forgetting.

4.2. Explainability and Interpretability Analysis

To demystify decision-making and ensure models avoided spurious memorization, LIME and SHAP were applied.

Authentic News Indicators: Both frameworks consistently attributed negative weights (predicting "Real") to factual terminology. Tokens like "said," "according to", “রিপোর্ট” (report - Bangla), and "के अनुसार" (according to - Hindi) universally marked authentic journalism.

Deceptive News Indicators: Conversely, models heavily relied on stylistic markers of sensationalism to classify "Fake" content. Capitalized urgency markers ("SHOCKING", "BREAKING"), local equivalents ("চাঞ্চল্যকর" in Bangla, "ब्रेकिंग" in Hindi, "ഞെട്ടിക്കുന്ന" in Malayalam), and English-derived code-mixed terms (e.g., "viral") drove deception classification. This confirms transformers successfully learned genuine cross-lingual stylistic cues.

5. Discussion and Limitations

This study advances multilingual disinformation detection via rigorous zero-shot and few-shot evaluations on real-world Indic datasets, achieving state-of-the-art few-shot performance (0.945 Macro-F1 on Hindi) with minimal annotation. Moving beyond broad surveys, this work introduces comprehensive explainability analysis across distinct Indic scripts, revealing underexplored cross-lingual stylistic consistencies. Despite these contributions, limitations define pathways for future research. First, traditional baselines were restricted to English and Bangla due to zero-shot constraints. Second, varying target dataset scales may have impacted few-shot stability for Malayalam and Tamil. Finally, constrained to textual features and binary classification, future architectures must integrate multimodal signals (images, captions) and transition toward multi-class frameworks capable of discerning varying shades of disinformation and their propagation dynamics.

6. Conclusion

This study systematically investigated multilingual transformer models for zero-shot and few-shot cross-lingual disinformation detection from English to four low-resource Indic languages (Bangla, Hindi, Malayalam, and Tamil). The results demonstrate that while large-scale models generalize significantly without target-language supervision, highly constrained fine-tuning yields profound performance enhancements. In zero-shot settings, XLM-RoBERTa (large) emerged as the most robust model (peak Macro-F1 of 0.726 on Bangla), underscoring the value of massive, heterogeneous pre-training. MuRIL also demonstrated competitive generalization, particularly for Hindi, reflecting the benefits of Indic-specific pre-training. Conversely, an ablation study utilizing traditional TF-IDF baselines failed entirely in cross-lingual settings, establishing deep contextual embeddings as a prerequisite for multilingual detection. Furthermore, few-shot adaptation with merely 200 labeled samples per language substantially bridged the cross-lingual performance gap, with MuRIL attaining near-optimal Macro-F1 scores of 0.945 for Hindi and 0.934 for Tamil. Finally, the pioneering application of post-hoc explainability frameworks (LIME and SHAP) confirmed these transformers successfully identify universal, cross-lingual stylistic deception cues—such as sensationalism, urgency markers, and code-mixing—rather than relying on spurious, domain-specific memorization.

Future Work

While establishing a robust foundation for cross-lingual disinformation detection, critical avenues remain for future exploration:

- **Expanded Linguistic Coverage:** Extending evaluations to a broader spectrum of low-resource Indic languages (e.g., Assamese, Odia, Punjabi, Telugu) is vital for understanding intra-family transfer dynamics and prioritizing data collection.
- **Multimodal Integration:** Integrating visual content analysis and caption-aware modeling will mitigate reliance on textual stylometric and improve resilience against sophisticated multimodal synthetic media.
- **Advanced Adaptation & Real-Time Architectures:** Investigating parameter-efficient fine-tuning (e.g., LoRA) minimizes annotation overhead. Furthermore, integrating these models into real-time streaming pipelines with active learning and fact-checking cross-referencing is crucial for moderating live social media environments.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] World Economic Forum. Global Risks Report 2025. Published January 15, 2025. 2025. URL: https://www.weforum.org/publications/global-risks-report2025/.

[2] Gordon Pennycook et al. "Shifting attention to accuracy can reduce misinformation online." In: Nature 592.7855 (2021), pp. 590–595. DOI: 10.1038/s41586-021-03344-2. URL: https://www.nature.com/articles/s41586-021-03344-2.

[3] Soroush Vosoughi, Deb Roy, and Sinan Aral. "The spread of true and false news online." In: Science 359.6380 (2018), pp. 1146–1151. DOI: 10.1126/science.aap9559. URL: https://www.science.org/doi/10.1126/science.aap9559.

[4] Gordon Pennycook and David G Rand. "Fighting misinformation on social media using crowdsourced judgments of news source quality." In: Proceedings of the National Academy of Sciences 116.7 (2019), pp. 2521–2526. DOI: 10.1073/pnas.1806781116. URL: https://www.pnas.org/doi/10.1073/pnas.1806781116.

[5] Stephan Lewandowsky, Ullrich KH Ecker, and John Cook. "Beyond misinformation: Understanding and coping with the "post-truth" world." In: Journal of Applied Research in Memory and Cognition 6.4 (2017), pp. 353–369. DOI: 10.1016/j.jarmac.2017.07.008. URL: https://www.sciencedirect.com/science/article/pii/S2211368117300713.

[6] David MJ Lazer et al. "The science of fake news." In: Science 359.6380 (2018), pp. 1094–1096. DOI: 10.1126/science.aao2998. URL: https://www.science.org/doi/10.1126/science.aao2998.

[7] Hunt Allcott and Matthew Gentzkow. "Social media and fake news in the 2016 election." In: Journal of Economic Perspectives 31.2 (2017), pp. 211–236. DOI: 10.1257/jep.31.2.211. URL: https://www.aeaweb.org/articles?id=10.1257/jep.31.2.211.

[8] Fabio Pierri and Stefano Ceri. "The COVID-19 social media infodemic." In: Scientific Reports 10.1 (2020), p. 16598. DOI: 10.1038/s41598-020-73510-5. URL: https://www.nature.com/articles/s41598-020-73510-5.

[9] Md Saiful Islam et al. "COVID-19–related infodemic and its impact on public health: A global social media analysis." In: The American Journal of Tropical Medicine and Hygiene 103.4 (2020), pp. 1621–1629. DOI: 10.4269/ajtmh.20-0812. URL: https://www.ajtmh.org/view/journals/tpmd/103/4/article-p1621.xml.

[10] Aengus Bridgman et al. "The causes and consequences of COVID-19 misperceptions: Understanding the role of news and social media." In: Harvard Kennedy School Misinformation Review 2.3 (2021). DOI: 10.37016/mr-2020-71. URL: https://www.google.com/search?q=https://misinforeview.hks.harvard.edu/article/the-causes-and-consequencesof-covid-19-misperceptions-understanding-the-role-of-news-andsocial-media/.

[11] Konstantinos I. Roumeliotis, Nikolaos D. Tselikas, and Dimitrios K. Nasiopoulos. "Fake news detection and classification: A comparative study of convolutional neural networks, large language models, and natural language processing models." In: Future Internet 17.1 (2025), p. 28. DOI: 10.3390/fi17010028. URL: https://www.mdpi.com/1999-5903/17/1/28.

[12] Esther Irawati Setiawan et al. "An image and text-based fake news detection with transfer learning." In: PLOS ONE 20.6 (2025), e0324394. DOI: 10.1371/journal.pone.0324394. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0324394.

[13] JH Yoon et al. "Triple-modality interaction for deepfake detection in zero-shot settings." In: Neurocomputing 475 (2025), pp. 1–10. DOI: 10.1016/j.neucom.2024.127456. URL: https://www.sciencedirect.com/science/article/abs/pii/S0925231224002021.

[14] Jacob Devlin et al. "BERT: Pre-training of deep bidirectional transformers for language understanding." In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (2019), pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423/.

[15] Yinhan Liu et al. "RoBERTa: A robustly optimized BERT pretraining approach." In: arXiv preprint arXiv:1907.11692 (2019). DOI: 10.48550/arXiv.1907.11692. URL: https://arxiv.org/abs/1907.11692.

[16] Salar Mohtaj et al. "Monolingual and Multilingual Misinformation Detection for LowResource Languages: A Comprehensive Survey." In: arXiv preprint arXiv:2410.18390 (2024). DOI: 10.48550/arXiv.2410.18390. URL: https://arxiv.org/abs/2410.18390.

[17] J Alghamdi et al. "Fake news detection in low-resource languages: A novel and efficient multilingual approach." In: Knowledge-Based Systems 297 (2024), p. 111928. DOI: 10.1016/j.knosys.2024.111928. URL: https://www.sciencedirect.com/science/article/pii/S0950705124005185.

[18] [18] Divyanshu Kakwani et al. "IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages." In: (2020). URL: https://aclanthology.org/2020.findings-emnlp.445/.

[19] [19] Pratik Joshi et al. "The state and fate of linguistic diversity and inclusion in the NLP world." In: arXiv preprint arXiv:2004.09095 (2020). DOI: 10.48550/arXiv.2004.09095. URL: https://arxiv.org/abs/2004.09095.

[20] Yuyang Yang et al. "Unsupervised fake news detection on social media: A generative approach." In: Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019), pp. 5644–5651. DOI: 10.1609/aaai.v33i01.33015644. URL: https://ojs.aaai.org/index.php/AAAI/article/view/4390.

[21] Feng Qian et al. "Neural user response generator: Fake news detection with collective user intelligence." In: IJCAI (2018), pp. 3834–3840. DOI: 10.24963/ijcai.2018/533. URL: https://www.ijcai.org/proceedings/2018/533.

[22] Iftitahu Ni'mah et al. "A simple contrastive embedding framework for low-resource fake news detection." In: Neural Computing and Applications (2025). DOI: 10.1007/s00521-025-11467-0. URL: https://link.springer.com/article/10.1007/s00521-025-11467-0.

[23] Alexis Conneau et al. "Unsupervised Cross-lingual Representation Learning at Scale." In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020), pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. URL: https://aclanthology.org/2020.acl-main.747.

[24] Simran Khanuja et al. "MuRIL: Multilingual Representations for Indian Languages." In: (2021). URL: https://arxiv.org/abs/2103.10730.

[25] Nils Reimers and Iryna Gurevych. "Making monolingual sentence embeddings multilingual using knowledge distillation." In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (2020), pp. 4512–4525. DOI: 10.18653/v1/2020.emnlp-main.365. URL: https://aclanthology.org/2020.emnlp-main.365/.

[26] Cynthia Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." In: Nature Machine Intelligence 1.5 (2019), pp. 206–215. DOI: 10.1038/s42256-019-0048-x. URL: https://www.nature.com/articles/s42256-019-0048-x.

[27] Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences." In: Artificial Intelligence 267 (2019), pp. 1–38. DOI: 10.1016/j.artint.2018.07.007. URL: https://www.sciencedirect.com/science/article/pii/S0004370218305988.

[28] Don Fallis. "What is disinformation?" In: Library Trends 63.3 (2015), pp. 401–426.

[29] Edson C Tandoc Jr, Zheng Wei Lim, and Rich Ling. "Defining 'fake news': A typology of scholarly definitions." In: Digital Journalism 7.2 (2019), pp. 137–153.

[30] Claire Wardle. "Fake news. It's complicated." In: First Draft (2017).

[31] Joshua A Tucker et al. "Social media, political polarization, and political disinformation: A review of the scientific literature." In: Hewlett Foundation (2018).

[32] Adam Badawy, Emilio Ferrara, and Kristina Lerman. "Characterizing the 2016 Russian IRA influence campaign." In: Social Network Analysis and Mining (2019).

[33] Shubhi Bansal et al. "MMCFND: Multimodal Multilingual Caption-Aware Fake News Detection for Low-Resource Indic Languages." In: arXiv preprint arXiv:2410.10407 (2024).

[34] Thomas Mandl et al. "Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages." In: Proceedings of the 11th Forum for Information Retrieval Evaluation (2019), pp. 14–17.

[35] Xinyi Zhou and Reza Zafarani. "Fake News: A Survey of Research, Detection Methods, and Opportunities." In: ACM Computing Surveys 53.5 (2020).

[36] Kai Shu et al. "Fake news detection on social media: A data mining perspective." In: ACM SIGKDD Explorations Newsletter 19.1 (2017), pp. 22–36.

[37] Alessandro Bondielli and Francesco Marcelloni. "A survey on fake news and rumour detection techniques." In: Information Sciences 497 (2019), pp. 38–55.

[38] Victoria L Rubin and Tatiana Lukoianova. "Truth and deception at the rhetorical structure level." In: Journal of the Association for Information Science and Technology 66.4 (2015), pp. 763–774.

[39] Martin Potthast et al. "A stylometric inquiry into hyperpartisan and fake news." In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (2017), pp. 231–240.

[40] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. "Information credibility on twitter." In: Proceedings of the 20th international conference on World wide web (2011), pp. 675–684.

[41] Liang Wu and Huan Liu. "Tracing fake-news footprints: Characterizing social media messages by how they propagate." In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (2018), pp. 637–645.

[42] Abhinav Gupta et al. "An Emotion and Novelty-Aware Approach for Multilingual Multimodal Fake News Detection." In: Findings of the Association for Computational Linguistics: AACL-IJCNLP (2022), pp. 494–505.

[43] Kai Shu et al. "FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media." In: Big Data 8.3 (2020), pp. 171–188.

[44] Zichao Yang et al. "Hierarchical attention networks for document classification." In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics (2016), pp. 1480–1489.

[45] Yaqing Wang et al. "EANN: Event adversarial neural networks for multi-modal fake news detection." In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2018), pp. 849–857.

[46] Jiangshu Du et al. "Cross-lingual COVID-19 Fake News Detection." In: arXiv preprint arXiv:2110.06495 (2021).

[47] Ashish Vaswani et al. "Attention is all you need." In: Advances in Neural Information Processing Systems. Vol. 30 (2017).

[48] Mina Schütz et al. "AITF HSTPatCheckThat!2022 : Cross-lingual Fake News Detection with trained Transformer." In: Working Notes of CLEF 2022 (2022).

[49] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach." In: Multimedia Tools and Applications. Vol. 80 (2021), pp. 11765–11788.

[50] Telmo Pires, Eva Schlinger, and Dan Garrette. "How multilingual is multilingual BERT?" In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (2019), pp. 4996–5001.

[51] Sangdo Han. "Cross-lingual Transfer Learning for Fake News Detector in a LowResource Language." In: arXiv preprint arXiv:2208.12482 (2022).

[52] Kai Shu et al. "Detecting fake news with weak social supervision." In: IEEE Intelligent Systems 34.6 (2019), pp. 24–34.

[53] AB Athira, SD Madhu Kumar, and Anu Mary Chacko. "A systematic survey on explainable AI applied to fake news detection." In: Engineering Applications of Artificial Intelligence 122 (2023), p. 106087.

[54] A. Rananga et al. "Misinformation detection on online social networks using explainable AI." In: Information Sciences 610 (2023), pp. 1–15.

[55] Dylan Slack et al. "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods." In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (2020), pp. 180–186.