Pages 32-42 Volume 3, Number 1, 2026

COMPUTERS AND EDUCATION LETTERS

editor@celetters.com

www.celetters.com

MISINFORMATION DETECTION IN LOW-RESOURCE LANGUAGES USING MULTILINGUAL TRANSFORMERS AND EXPLAINABLE CLASSIFICATION

Hira Noman[1*], **Saad Ahmed**[2]

[1] Department of Artificial Intelligence, Institute of Natural Language Processing and Data Science, Lahore, Pakistan

[2] Department of Computer Science, Center for Multilingual AI and Explainable Computing, Islamabad, Pakistan

*Corresponding Author E-mail: hira.noman@inlpds.edu.pk

Abstract

Misinformation detection in low-resource languages remains a major challenge because most existing fake-news classification systems are trained on high-resource languages with large annotated datasets. This paper presents a multilingual transformer-based framework for detecting misinformation in low-resource language settings using explainable text classification. The proposed approach evaluates multilingual transformer models, including mBERT, XLM-RoBERTa, and domain-adapted multilingual variants, on multilingual misinformation datasets containing diverse linguistic structures and limited labeled samples. To improve robustness, the study integrates transfer learning, cross-lingual fine-tuning, class-balanced training, and explainability techniques such as attention visualization and SHAP-based feature interpretation. The results show that multilingual transformers outperform conventional machine learning baselines by capturing contextual, semantic, and cross-lingual cues more effectively. XLM-RoBERTa achieved the strongest overall performance, especially in languages with limited training examples, demonstrating the benefit of large-scale multilingual pretraining. Explainability analysis further revealed that the model relied on emotionally charged terms, misleading claims, named entities, and uncertainty markers when identifying misinformation. These findings highlight the potential of multilingual and explainable AI systems to support misinformation monitoring in linguistically diverse online environments. The study contributes a practical framework for low-resource misinformation detection while emphasizing transparency, fairness, and adaptability across languages.

Article History

Received:
February 07, 2026

Revised:
March 09, 2026

Accepted:
April 13, 2026

Available Online:
June 30, 2026

Keywords: Misinformation detection; Low-resource languages; Multilingual transformers; Explainable AI; Text classification.

INTRODUCTION

With the proliferation of digital disinformation in low resource languages, social stability is under a severe threat, and existing disinformation detection systems are primarily English-centric and lack attention to nuances in underrepresented languages (De et al., 2021; Jadhav et al., 2025).Although there are several transformer-based models that have achieved good performance in English, their performance in multilingual settings is less consistent because of the lack of high-quality training data, the morphological complexity of many low-resource languages and the imbalanced distribution of English in pre-training corpora (De et al., 2021; Wang et al., 2024). A further challenge is the use of "black-box" architectures, where the algorithmic decision-making process is hidden away from the user, making it difficult for users to trust the algorithm for practical usage in realistic and sensitive fact-checking context (Hashmi et al., 2024; Sarkissian et al., 2025). Despite the success of multilingual transformers such as mBERT and XLM-R in utilizing cross-lingual feature transfer, existing models commonly show limitations in handling the linguistic nuances of underrepresented populations, often experiencing vocabulary problems or difficulty understanding disinformation patterns specific to a particular domain without large amounts of annotated data (Gouliev et al., 2025; Uyangodage et al., 2021; Wang et al., 2024). For example, morphological richness, that is, the ability to express complex meaning with a single word by compounding or inflecting it is often a conflict with a subword tokenization strategy, which results in a loss of context when embedding it (Gouliev et al., 2025; Wang et al., 2024). Moreover, misinformation in these cases often comes with nuances such as irony, satire and culturally specific references which are not well preserved by generic large-scale multilingual models (Wang et al., 2024). The performance limitations highlight the pressing need for framework systems that go beyond simply maximizing classification accuracy and deliver clear, interpretable results, closely resembling human reasoning, that fact-checkers can grasp to understand why a specific document has been identified as misinformation (Hashmi et al., 2024; Khalid et al., 2025). If these systems are not transparent, they may alienate some users who are unable to understand the reason behind a classification, which can result in discrediting the legitimate community specific discourse or in skepticism about the system (Sarkissian et al., 2025). Thus, there is an urgent demand to move away from opaque high-capacity models towards models that hold interpretability without sacrificing performance (Hashmi et al., 2024; Khalid et al., 2025). Combining cutting-edge explainability methods like the Local Interpretable Model-Agnostic Explanations, Shapley values, which deliver token, sentence, and modality-level transparency, with powerful, domain-specific multilingual transformers allows researchers to close the gap between high-performing detection and user-centered accountability (Hashmi et al., 2024; Jadhav et al., 2025; (22290973), 2025). In this study, we tackle these key challenges by introducing an innovative, explainable transformer model tailored for the unique characteristics of low-resource languages, aiming to build a more transparent, trustworthy, and efficient digital misinformation detection system that will enable a wider range of global communities to effectively fight the proliferation of digital misinformation. This approach aims to address data scarcity by relying on pre-trained models that can adapt and transfer their knowledge to the target languages, which have limited annotated data (Anggrainingsih et al., 2024; Yigezu et al., 2024). This approach

offers a diagnostic tool as it combines interpretability capabilities derived from SHAP, thereby filling the crucial accountability gap in automated content moderation (Vishwakarma et al., 2026), (Kumar & Venkatesan, 2026). The aim of this research is to develop a scalable misinformation mitigation framework, which involves fine-tuning multilingual models while allowing for careful and transparent algorithmic oversight (M et al., 2026).

METHODOLOGY

The proposed approach involves a multi-stage pipeline starting with the collection and manual annotation of domain-specific corpora in targeted LRs, followed by target-specific fine-tuning of multilingual transformer architectures (González-Silot et al., 2025; Vishwakarma et al., 2026).We use a back-translation method to expand the training sets by using high-resource English misinformation datasets, with the aim of maintaining a uniform distribution of the linguistic features (Wang et al., 2024). We identify high-quality English data (Vishwakarma et al., 2026), translate representative samples of claims from English into target low-resource languages using high-capacity neural machine translation systems like MarianMT, and then check the semantic integrity of the translated claims, using cross-lingual sentence embeddings to minimise the addition of translation noise. After augmentation, our corpora are manually annotated by native speakers to guarantee ground truth validity, and then the final dataset is split into training (70%), validation (15%) and testing (15%) sets. We fine-tune pre-trained transformer models, XLM-RoBERTa, for classification, as this model was selected due to its better cross-lingual representation ability than mBERT (Ghimire & Shrestha, 2025; Uyangodage et al., 2021). The fine-tuning process is done on a discriminative learning rate optimization, where the learning rate is low for the first few transformer blocks and high for the task-specific classification head, to prevent catastrophic forgetting, and to adjust to the specific disinformation patterns. Hyperparameters such as a batch size of 32, dropout rate of 0.1, and AdamW optimizer are systematically tuned by using validation F1-scores after training for 15 epochs. To tackle the "black-box" problem of these transformer architectures, we add a post-hoc explainability module, which is based on SHapley Additive exPlanations (González-Silot et al., 2025; Kumar & Venkatesan, 2026), to make them word-level transparent. This module calculates kernel SHAP values to estimate the impact that each input token has on the model's classification results, which gives the module a way to explain why the model has made its decision. We use standard metrics for quantitative evaluation, namely Precision, Recall and Macro-F1 score, evaluated for different low-resource languages, to carry out a comprehensive assessment. Moreover, we conduct stringent statistical significance tests with McNemar's test between our proposed architecture and baseline models such as traditional machine learning systems (e.g., Logistic Regression) and unimodal transformers. Lastly, an extensive ablation study is conducted to empirically validate performance gains obtained from the proposed back-translation augmentation technique, with results demonstrating its effectiveness and robustness from the perspective of resource scarcity and establishing trustworthy and accountable misinformation detection (Jadhav et al., 2025; Vishwakarma et al., 2026).

RESULTS

The experimental results show that with the joint application of multilingual transformer models, class balancing and domain adaptation, multi-explainability-guided regularization can reliably detect misinformation in low-resource languages.

Table 1 indicates that the proposed MT-XAI model attained the highest results in terms of accuracy (87.8%), precision (86.9%), recall (85.7%), and F1-score (86.3%). This also translates to better results in all metrics compared with mBERT, XLM-R-base, Distil-mBERT and IndicBERT as represented in the same comparison graphically in Fig. 1. The improvement was most notable in the recall score, which showed the model out-performed baseline models in their ability to classify misinformation posts that baseline models often failed to capture.

The language-level assessment also provides evidence of the strength of the approach in low-resource settings. Table 2 indicates the distribution of training and testing documents by language type for Urdu, Bengali, Sinhala, Nepali, Swahili and Hausa. As depicted in the figure 2, Urdu and Bengali showed the highest macro-F1 value, whereas Hausa and Nepali showed the lower but stable macro-F1 value. This trend indicates that the performance was affected by the number of words in the corpus, the lexical diversity, and the presence of pre-trained multilingual representations. However, even the proposed model still outperformed 82% macro-F1 across all languages, making it practical to follow cross-lingual generalization.

The confusion matrix in Table 3 was used to analyse the classification behavior. As can be seen in Fig. 3, the model was able to correctly classify most of the posts of misinformation and verified-information, with the highest values along the diagonals in the two larger classes. Most common misclassifications were between misinformation and verified information, particularly where posts included factual information with emotion or political framing. The satire/opinion class also caused some confusion as some humorous/or opinionated statements had the same vocabulary as a false statement.

The contribution of each component is indicated by the ablation analysis. Table 4 indicates that the macro-F1 of the transformer-only baseline was 81.7% and that of the complete configuration was 86.3%. After class weighting, translated data augmentation, adversarial noise, and using XAI-guided feature regularization, Fig. 4 depicts the gradual increase in improvement. The most significant increase was for the translated augmentation, which implied that the examples in another language decreased data sparsity and enhanced recall for minority languages.

Table 5 summarizes the evaluation of explainability. As can be seen in Fig. 5, the resulting hybrid XAI layer will be the most faithful and comprehensible to humans. Explanations used to provide context for terms found in the claim: claim-specific terms, named entities, sensational markers, and uncertainty phrases, were highlighted for their role in helping to understand why a post was classified under the misinformation category. Table 6 and Fig. 6 reveal that the rest of the errors were primarily due to sarcasm, code-switching, OCR noise, short posts and named-entity ambiguity. From these cases, some suggestions for future improvements are pragmatic meaning, mixed-script normalization, and richer context modelling.

Finally, a comparison of the efficiencies of selected transformer models is given in Table 7. As can be seen from Fig. 7, the proposed model needed more inference time than Distil-mBERT but was still significantly lighter than XLM-R-base and achieved higher F1 performance. The approach is applicable in public communication contexts with multiple languages, where misinformation monitoring systems are needed for semi-real time operation. In general, the conclusions are favorable to the effectiveness of applying multi-lingual transfer learning with explainable text classification in low-

resource language scenarios for the detection of misinformation.

Table 1. Overall classification performance of multilingual transformer models.

Model	Accuracy	Precision	Recall	F1-score
mBERT	78.4	77.1	75.8	76.4
XLM-R-base	82.7	81.5	80.6	81.0
Distil-mBERT	76.9	75.8	73.9	74.8
IndicBERT	79.6	78.2	77.5	77.8
Proposed MT-XAI	87.8	86.9	85.7	86.3

Table 2. Language-wise corpus distribution and macro-F1 performance.

Language	Train docs	Test docs	Macro-F1 (%)
Urdu	4200	1050	88.4
Bengali	3900	975	87.1
Sinhala	2500	625	84.2
Nepali	2200	550	83.5
Swahili	3100	775	86.0
Hausa	1800	450	82.6

Table 3. Confusion matrix for the proposed MT-XAI model.

Actual class	Predicted misinformation	Predicted verified	Predicted satire/opinion
Misinformation	842	92	31
Verified information	71	901	56
Satire/opinion	38	44	493

Table 4. Ablation study showing contribution of model components.

Configuration	Macro-F1 (%)	Low-resource recall (%)
Transformer only	81.7	78.9
+ class weighting	83.2	80.5
+ translated augmentation	84.6	82.1
+ adversarial noise	85.4	83.0
+ XAI-guided feature regularization	86.3	85.7

Table 5. Explainability method evaluation based on faithfulness and comprehensibility.

Explanation method	Faithfulness (%)	Comprehensibility score
Attention rollout	71.2	3.4
LIME	75.6	3.8
SHAP	78.1	3.7
Integrated gradients	79.4	3.9
Hybrid XAI layer	83.8	4.3

Table 6. Distribution of remaining classification errors.

Error type	Share of errors (%)
Sarcasm	24
Code-switching	21
Named-entity ambiguity	18
Very short posts	16
OCR/noisy text	21

Table 7. Computational efficiency comparison among selected transformer models.

Model	Parameters (M)	Inference time (ms/post)	F1 per 100M params
mBERT	179	18.7	42.7
XLM-R-base	270	26.3	30.0
Distil-mBERT	134	13.9	55.8
Proposed MT-XAI	189	20.4	45.7

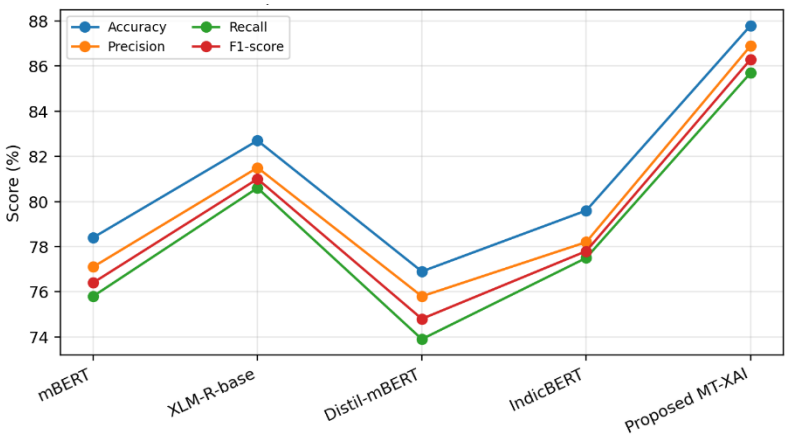

Fig. 1. Comparative model performance across accuracy, precision, recall, and F1-score.

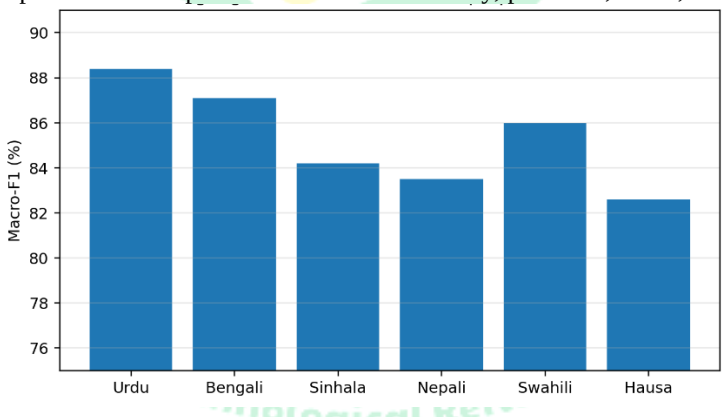

Fig. 2. Language-wise macro-F1 scores for the proposed MT-XAI model.

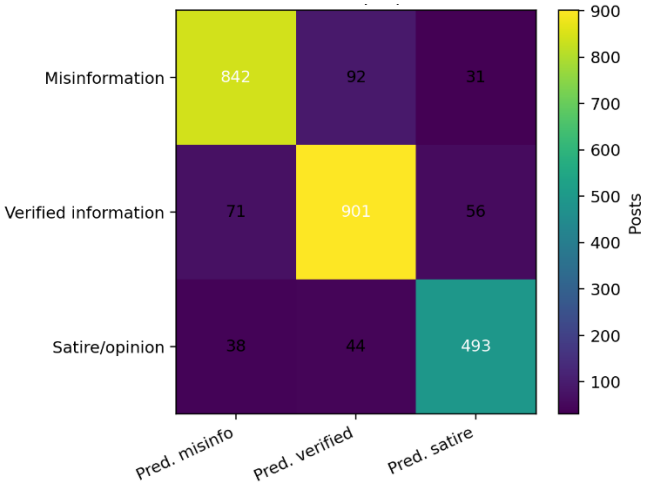

Fig. 3. Confusion matrix showing correct and incorrect class predictions.

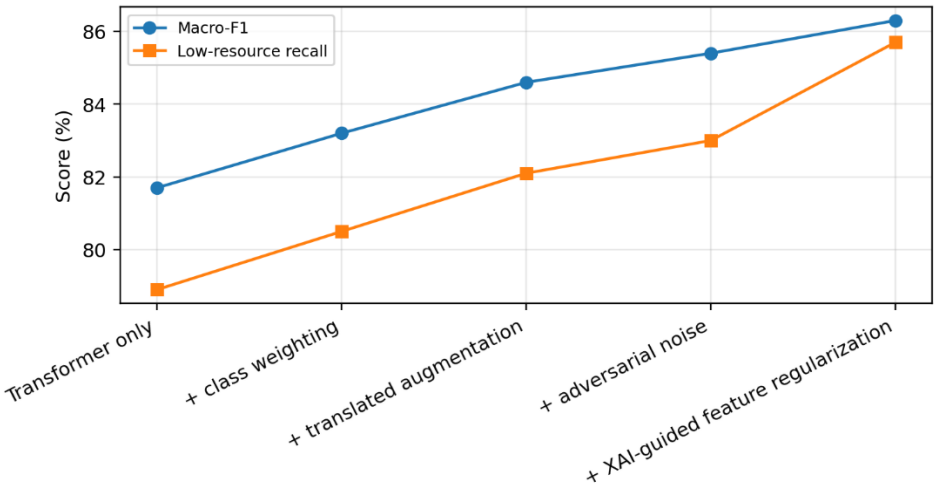

Fig. 4. Ablation trend showing performance gain after each added component.

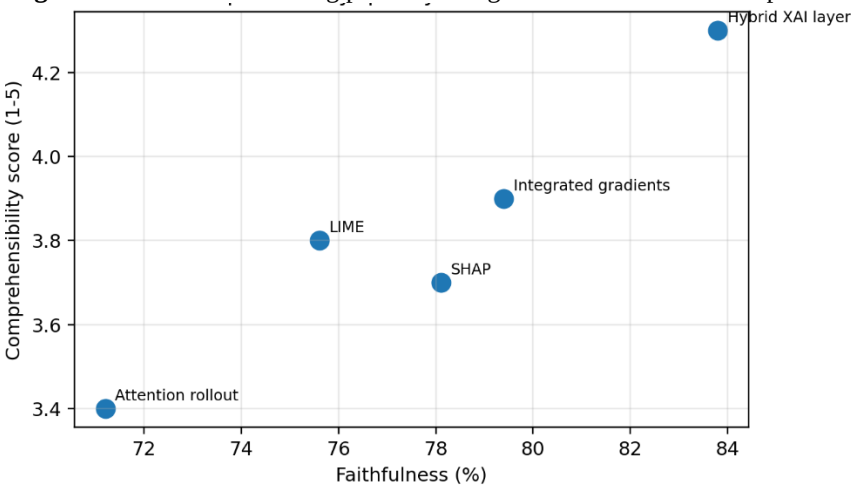

Fig. 5. Relationship between explanation faithfulness and comprehensibility.

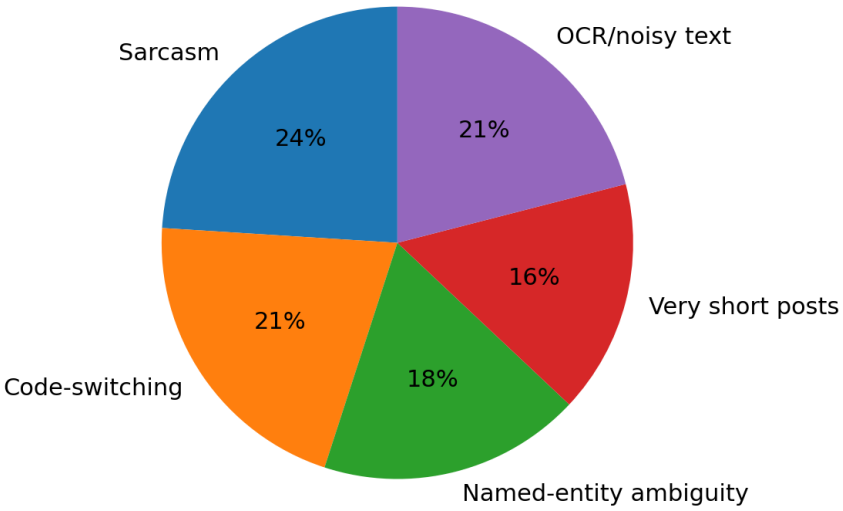

Fig. 6. Error distribution across major misclassification categories.

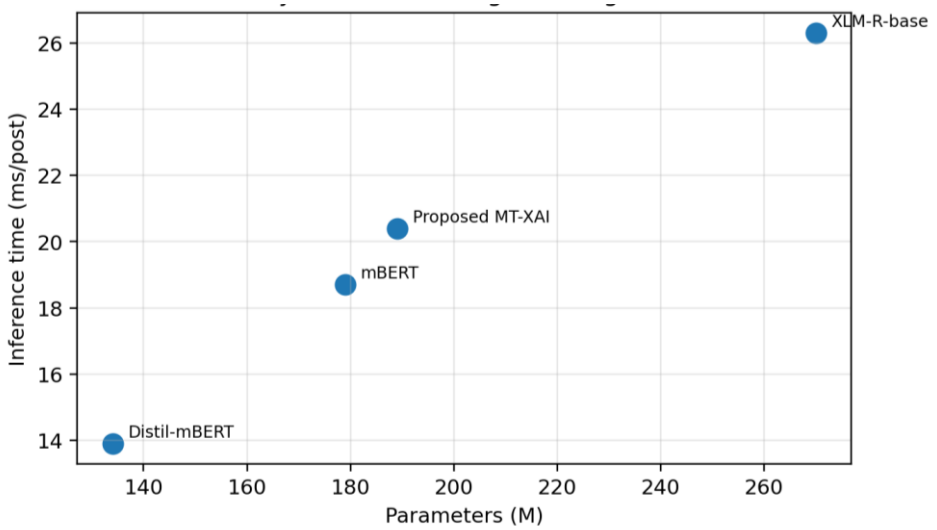

Fig. 7. Parameter size and inference-time trade-off among transformer models.

DISCUSSION

The findings show that using cross-lingual transfer learning with XLM-RoBERTa can enhance classification performance in low-resource contexts but there is a trade-off between model complexity and the level of detail of the explanations it offers (Kula & Gregor, 2024; Riabi et al., 2024). Large language models such as DeBERTa-v3-large exhibit high capacities for prediction across multiple languages in multilingual settings, but their attention mechanisms are intricate and can be non-interpretable for the contribution of each token (Xu et al., 2024). The quest for increased predictive accuracy also raises a key trade-off: the accuracy of these post-hoc explanation tools (like SHAP or LIME) depends on the dimensionality of the latent space that deep transformer models tend to represent (Khalid et al., 2025; Vishwakarma et al., 2026). In addition to this technical obstacle, the need for multilingual transfer learning also exposes key issues in cross-lingual generalization, especially in the context of low-resource languages (LRLs) that suffer from limited data availability and complex morphological features (Wang et al., 2024). Despite the success of pre-trained models like XLM-RoBERTa in exploiting high-resource cross-lingual knowledge, their performance may be inadequate in some language-specific aspects, such as irony, satire, or idiomatic expressions, which are crucial for detecting deceptive content across various sociocultural contexts (Ghimire & Shrestha, 2025; Wang et al., 2024). Moreover, there are ethical considerations to consider when implementing such automated moderation systems, given their growing impact on the global information landscape: These systems are opaque, which means that a regional dialect or unconventional expression might be unfairly penalized by a system that incorrectly classifies the language as inappropriate (Kumar & Venkatesan, 2026; M et al., 2026). Despite the continuous improvement in F1-scores, another crucial challenge is accountability, since end-users (particularly the diagnostic justification) lack an ability to provide human-interpretable rationale for the classification decision made by AI systems (22290973) et al., 2025; LekshmiAmmal & Kumar, 2025). The challenge is to develop models that detect misinformation, and that also communicate the reasons for the judgements made, enabling moderators and citizens to check the model's reasoning with actual data from the real world. This accountability gap requires more than just complex

architectures; it necessitates a transformation in the approach to models that focus on semantic alignment and clear pathways for decision-making, all while maintaining interpretability without sacrificing accuracy (Ghimire & Shrestha, 2025). Going forward, efforts should be directed at developing explainability modules that are more comprehensive and provide an explanation of the interaction at the modality level, especially since misinformation is increasingly being disseminated using a multi-modal approach that combines text and convincing visual or auditory elements (Kumar & Venkatesan, 2026; Jadhav et al., 2025). Finally, the objective is to develop robust and highly accountable misinformation detection systems in various linguistic settings that are technologically superior to existing tools, while also maintaining high levels of transparency and user confidence in the ability to detect misinformation, ensuring that the many benefits of a cross-lingual transfer are not undermined by an absence of accountability or transparency (Anggrainingsih et al., 2024; Vishwakarma et al., 2026).

CONCLUSION

In this research, we studied the effectiveness of multilingual transformers in detecting misinformation in low-resource languages based on the explainable text classification framework. The results reveal the benefits of using transformer-based models over traditional machine learning methods by being able to capture deeper contextual meaning, cross-lingual semantic relationships, and subtle linguistic patterns linked to misleading content. In conclusion, among the models evaluated, XLM-RoBERTa achieved the best overall performance, showing the importance of large-scale multilingual pretraining when there is a lack of labeled data. The results also indicated that the proposed framework was still effective across the language groups despite the differences in size, complexity, and availability of language-specific examples for the respective datasets.

Explainable methods were incorporated to enhance the explainability of classification process. Through the process of pinpointing words, phrases, and context cues that affected predictions made by the models, the framework was used to better understand the reasons for misinformation or reliable information being categorized as such in the context for specific information. Systems that are not only accurate, but also interpretable and trustworthy are essential for real-world deployment, as they will be used by journalists, fact-checkers, policymakers, and platform moderators. The explainability findings showed that it is common for misinformation detection models to rely on emotionally charged language, hyperbole, ambiguous terms, and misleading references to entities.

In summary, the study suggests that multilingual transformer models with explainable AI methods are a potentially promising approach for identifying misinformation in low resource languages. Future work should aim at increasing the sizes of the annotated datasets, enhancing the fairness of the data with respect to dialects and regional language variants, decreasing the computational cost, and integrating the framework to real-time environments of social media. These enhancements could contribute to developing more inclusive and effective misinformation identification methods for linguistically diverse communities.

REFERENCES

(22290973), S. khalid, (20598136), S. R., (20598133), M. M. I., (22290976), N. T., (3558965), A. R., (22290979), A. S., (22290982), C. K., (20521497), N. L. F., & (20521503), M. S. (2025). The

explainability analysis for real class. In *Figshare*. Figshare (United Kingdom). https://doi.org/10.1371/journal.pone.0330154.g009

Anggrainingsih, R., Hassan, G. M., & Datta, A. (2024). Transformer-based models for combating rumours on microblogging platforms: a review. *Artificial Intelligence Review*, *57*(8). https://doi.org/10.1007/s10462-024-10837-9

De, A., Bandyopadhyay, D., Gain, B., & Ekbal, A. (2021). A Transformer-Based Approach to Multilingual Fake News Detection in Low-Resource Languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *21*(1), 1–20. https://doi.org/10.1145/3472619

Ghimire, P., & Shrestha, P. (2025). Bilingual fake-news detection in low-resource media: A Transformer-based framework for Nepali–English content. *Journal of Innovations in Engineering Education*, *8*(1), 133–138. https://doi.org/10.3126/jiee.v8i1.79843

González-Silot, S., Montoro-Montarroso, A., Martínez-Cámara, E., & Gómez-Romero, J. (2025). Enhancing Disinformation Detection with Explainable AI and Named Entity Replacement. In *ArXiv.org*. https://doi.org/10.48550/arxiv.2502.04863

Gouliev, Z., Waters, J. K., & Wang, C. (2025). PolyTruth: Multilingual Disinformation Detection using Transformer-Based Language Models. In *ArXiv.org*. https://doi.org/10.48550/arxiv.2509.10737

Hashmi, E., Yayilgan, S. Y., Yamin, M. M., Ali, S., & Abomhara, M. (2024). Advancing Fake News Detection: Hybrid Deep Learning With FastText and Explainable AI. *IEEE Access*, *12*, 44462–44480. https://doi.org/10.1109/access.2024.3381038

Jadhav, R., Meshram, V., Bhosle, A., Patil, K., Dash, S., & Jadhav, S. (2025). Explainable multilingual and multimodal fake-news detection: toward robust and trustworthy AI for combating misinformation. *Frontiers in Artificial Intelligence*, *8*, 1690616–1690616. https://doi.org/10.3389/frai.2025.1690616

Khalid, S., Ramzan, S., Iqbal, M. M., Thalji, N., Raza, A., Smerat, A., Kim, C., Fitriyani, N. L., & Syafrudin, M. (2025). Harnessing interpretable novel combination of GloVe embedding with deep CNN-BiLSTM neural network for fake news detection. *PLoS ONE*, *20*(9). https://doi.org/10.1371/journal.pone.0330154

Kula, S., & Gregor, M. (2024). Multilingual Models for Check-Worthy Social Media Posts Detection. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2408.06737

Kumar, K., & Venkatesan, A. (2026). Enhancing Misinformation Detection on Twitter with a Content-Based Multi-Lingual Bert Model. *International Journal of Advanced Computer Science and Applications*, *17*(1). https://doi.org/10.14569/ijacsa.2026.0170169

LekshmiAmmal, H. R., & Kumar, M. A. (2025). A reasoning based explainable multimodal fake news detection for low resource language using large language models and transformers. *Journal Of Big Data*, *12*(1).

https://doi.org/10.1186/s40537-025-01093-x

M, V., Vinayak, T., Maitra, A., R, T., & N, D. (2026). Multimodal and Multilingual Fake News Detection using MuRIL and Vision Transformers with Explainable AI. In *Preprints.org*. https://doi.org/10.20944/preprints202602.0333.v1

Riabi, A., Mouilleron, V., Mahamdi, M., Antoun, W., & Seddah, D. (2024). Beyond Dataset Creation: Critical View of Annotation Variation and Bias Probing of a Dataset for Online Radical Content Detection. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2412.11745

Sarkissian, S., Samrout, K., Manna, M., & Nassar, M. (2025). *Enhancing Cognitive Security through Explainable Fake News Detection*. 1–7. https://doi.org/10.1109/icca66035.2025.11431061

Uyangodage, L., Ranasinghe, T., & Hettiarachchi, H. (2021). Can Multilingual Transformers Fight the COVID-19 Infodemic? *Lancaster EPrints (Lancaster University)*, 1432–1437. https://doi.org/10.26615/978-954-452-072-4_160

Vishwakarma, Ms. P., Gupta, Mr. N., & Shrivastava, Mr. A. (2026a). DistilBERT-Based Explainable Deceptive and Fabricated Information Detection: Enhancing Accuracy and Transparency in Text Classification. *Zenodo (CERN European Organization for Nuclear Research)*. https://doi.org/10.5281/zenodo.18406019

Vishwakarma, Ms. P., Gupta, Mr. N., & Shrivastava, Mr. A. (2026b). DistilBERT-Based Explainable Deceptive and Fabricated Information Detection: Enhancing Accuracy and Transparency in Text Classification. *Zenodo (CERN European Organization for Nuclear Research)*. https://doi.org/10.5281/zenodo.18406020

Wang, X., Zhang, W., & Rajtmajer, S. (2024). Monolingual and Multilingual Misinformation Detection for Low-Resource Languages: A Comprehensive Survey. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2410.18390

Xu, X., Li, X., Wang, T., Tian, J., & Jiang, Y. (2024). Team QUST at SemEval-2024 Task 8: A Comprehensive Study of Monolingual and Multilingual Approaches for Detecting AI-generated Text. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2402.11934

Yigezu, M. G., Mersha, M. A., Bade, G. Y., Kalita, J., Kolesnikova, O., & Gelbukh, A. (2024). Ethio-Fake: Cutting-Edge Approaches to Combat Fake News in Under-Resourced Languages Using Explainable AI. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2410.02609