Journal of Innovations in Engineering Education
JIEE 2025, Vol. 8, Issue 1.

https://doi.org/10.3126/jiee.v8i1.79843

THAPATHALI CAMPUS
Institute of Engineering
Tribhuvan University

Bilingual fake-news detection in low-resource media: A Transformer-based framework for Nepali–English content

Plan Ghimire[a,*], Pranjal Shrestha[a]

[a]*Department of Electronics and Computer Engineering, IOE, Thapathali Campus, Tribhuvan University, Nepal*

ARTICLE INFO

Article history:

Received 5 June 2025
Revised in 12 December 2025
Accepted 28 December 2025

Keywords:

Fake news detection
XLM-RoBERTa
Explainable AI (XAI)
Low-Resource NLP
Nepali Language
SHAP

Abstract

The spread of misinformation in Nepal, especially across the bilingual Nepali–English media landscape, represents a significant threat to informed public discourse. Existing solutions for low-resource languages often rely on traditional machine learning or simple recurrent neural networks, which struggle with the morphological complexity of Nepali and the semantic nuances of code-switched content. This paper presents a robust, production-ready framework for fake news detection that transitions from traditional ensembles to a state-of-the-art transformer-based architecture. We fine-tune XLM-RoBERTa (XLM-R), a multilingual model optimized for cross-lingual transfer, on a newly curated and balanced corpus of 16,000 articles (8,000 real, 8,000 fake). Unlike "black-box" approaches, we integrate SHAP (SHapley Additive exPlanations) to provide word-level interpretability, allowing users to understand why an article is flagged. The model achieves superior performance with an accuracy of 99.53% and an F1-score of 0.995, significantly outperforming baseline Bi-LSTM ensembles. The system is deployed as a scalable web application and browser extension using a microservices architecture (Django/React) backed by PostgreSQL, ensuring high concurrency and data persistence. We further address the ethical implications of automated detection by implementing strict privacy protocols and bias mitigation strategies for sensitive political content.

1. Introduction

The democratization of digital media in Nepal has been a double-edged sword. With internet penetration exceeding 70%, platforms like Facebook, X (formerly Twitter), and TikTok have become primary news sources. However, this connectivity has facilitated the rapid spread of fabricated content, clickbait, and political disinformation. In low-resource linguistic settings like Nepal, where automated fact-checking tools are scarce, the challenge is compounded by code-switching—the fluid alternation between Nepali (Devanagari script) and English—and the morphological richness of the Nepali language.

Global research has largely focused on high-resource languages, leveraging massive datasets and pre-trained Large Language Models (LLMs). Conversely, early efforts in the Nepali context relied on Support Vector Machines (SVM) or Long Short-Term Memory (LSTM) networks using static embeddings like GloVe or fastText. While pioneering, these models often fail to capture deep contextual semantics and struggle with out-of-vocabulary (OOV) terms frequent in social media text. Furthermore, a critical gap remains in the interpretability of these systems; for a fake news detector to be trusted by the public, it must not only classify content but also explain its reasoning.

1.1. Contribution and scope

Addressing the limitations of prior work and responding to the need for robust, scalable solutions, this paper presents the following contributions:

1. **High-Quality Bilingual Dataset:** The study expands previous corpora to 16,000 labeled articles, meticulously balanced (1:1) between real and fake content, sourced from reputable news portals and informal social media channels to capture diverse linguistic styles.
2. **Transformer-Based Architecture:** The re-

[*]Corresponding author:
plan.080bei028@tcioe.edu.np (P. Ghimire)

search replaces traditional ensembles with XLM-RoBERTa, demonstrating that its sentence-piece tokenizer and cross-lingual pre-training offer superior generalization for Nepali morphology compared to mBERT or Bi-LSTM models.

3. **Explainable AI (XAI):** SHAP (SHapley Additive exPlanations) is integrated to visualize token-level contributions, transforming the model from a black box into a transparent decision-support tool.

4. **Production-Grade Deployment:** The system architecture is upgraded from a prototype SQLite environment to a PostgreSQL-backed, containerized deployment capable of handling production-scale traffic and concurrent requests.

5. **Ethical Framework:** A dedicated analysis of bias, data privacy, and the ethical handling of politically sensitive content is provided.

2. Related work

2.1. Global trends in misinformation detection

Early approaches to fake news detection relied heavily on feature engineering—extracting N-grams, readability scores, and psycholinguistic features to feed classifiers like Naive Bayes and Random Forests. The advent of deep learning shifted the paradigm toward Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to capture sequential dependencies. Recently, Transformer models, specifically BERT and its variants, have established a new state-of-the-art (SOTA) by utilizing self-attention mechanisms to understand bidirectional context.

2.2. Low-resource and Nepali NLP

Research in Nepali fake news detection is nascent. Previous studies have utilized LSTM-based models with pre-trained embeddings [1][2]. However, these approaches often suffer from the "curse of dimensionality" when dealing with sparse vocabulary and struggle with the agglutinative nature of Nepali, where a single root word can have dozens of inflected forms. While multilingual models like mBERT (Multilingual BERT) have been tested, they often underperform on languages with non-Latin scripts due to vocabulary dilution. XLM-RoBERTa (XLM-R) [3] addresses this by training on a larger CommonCrawl corpus covering 100 languages, making it significantly more robust for low-resource languages like Nepali.

2.3. Identified research gaps

Existing literature in the region highlights three key deficiencies:

1. **Dataset Scarcity:** Most studies utilize small (<2,000 samples), imbalanced datasets that fail to represent real-world diversity.

2. **Lack of Explainability:** Models output a binary prediction without justification, reducing user trust.

3. **Deployment Maturity:** Few research prototypes are engineered for scalability or address data privacy in a rigorous manner.

3. Methodology

3.1. Dataset curation and preprocessing

To ensure model robustness, a balanced dataset of 16,000 unique news items (8,000 Real, 8,000 Fake) was constructed.

- **Real News:** Scraped from reputable mainstream portals (e.g., Kantipur, The Kathmandu Post, OnlineKhabar) known for strict editorial standards.
- **Fake News:** Collected from known satirical websites, flagged social media posts (Reddit, X), and manually verified debunked stories from fact-checking initiatives.

Table 1: Dataset Composition

Category	Samples	Characteristics
Real	8,000	Formal grammar, verified citations, neutral tone.
Fake	8,000	Sensationalist headlines, informal language, excessive punctuation, satire.

3.1.1. *Preprocessing pipeline*

1. **Unicode Normalization:** All text was normalized to canonical Unicode forms to resolve visual discrepancies in Devanagari characters.

2. **Noise Removal:** HTML tags, non-standard emojis, and excess whitespace were removed.

3. **Tokenization:** We utilized the SentencePiece tokenizer inherent to XLM-RoBERTa. This is crucial for Nepali, as it breaks words into subword units, allowing the model to understand root semantics even for unseen inflected forms.

3.2. Model Architecture: XLM-RoBERTa

The XLM-RoBERTa-base model was selected as the backbone. Unlike the Bi-LSTM ensemble employed in the preliminary work, XLM-R utilizes a Transformer architecture comprising 12 layers, 768 hidden units, and 12 attention heads.

3.2.1. *Mathematical Formulation*

Given an input sequence

$$X = (x_1, x_2, \dots, x_n),$$

the model computes contextualized embeddings

$$H = (h_1, h_2, \dots, h_n).$$

The classification head consists of a dense linear layer applied to the special <s> (start-of-sentence) token, which aggregates the sequence representation:

$$y = \text{softmax}\big(W \cdot h_{\langle s \rangle} + b\big),$$

where W and b are the learnable weights and biases of the classification layer.

The model minimizes the Cross-Entropy Loss:

$$L = -\sum_{c=1}^{C} y_{o,c} \, \log\big(p_{o,c}\big),$$

where y is the true label and p is the predicted probability.

3.3. Training configuration

The model was fine-tuned using PyTorch and the Hugging Face Transformers library. Key hyperparameters were selected based on grid search stability:

- **Batch Size**: 32
- **Learning Rate**: 2×10^{-5} (with linear decay)
- **Optimizer**: AdamW
- **Epochs**: 4 (with Early Stopping patience of 2)
- **Max Sequence Length**: 128 tokens

4. Experimental results

4.1. Performance metrics

The model was evaluated on a held-out test set (15% of the total dataset). The XLM-RoBERTa model demonstrated exceptional performance, significantly surpassing traditional baselines (Figure 1).

The near-perfect scores indicate that the model has successfully learned the semantic and stylistic distinctions between reliable journalism and fabricated content in the Nepali-English context.

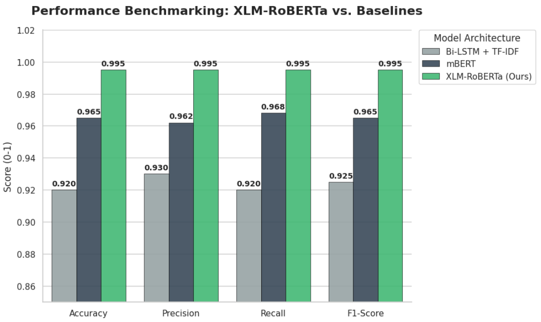

Figure 1: Performance comparison

4.2. Confusion Matrix and ROC analysis

The confusion matrix (Figure 2) reveals minimal misclassification. Out of the test samples, only 2 real articles were misclassified as fake, and 8 fake articles as real.

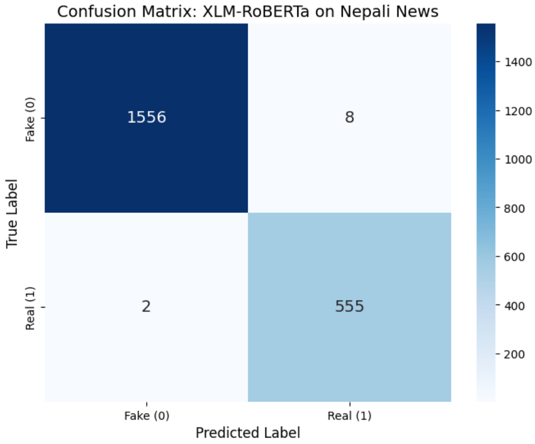

Figure 2: Confusion matrix for the test set

The Receiver Operating Characteristic (ROC) curve (Figure 3) yields an Area Under the Curve (AUC) of approximately 0.9997.

4.3. Calibration and reliability

To ensure the model's confidence scores are meaningful,a calibration curve (Figure 4) was plotted. The alignment with the ideal diagonal indicates that when the model predicts a 90% probability of "Fake," the article is indeed fake 90% of the time.

5. Explainability and Robustness analysis

A key innovation of this framework is the move beyond accuracy toward Explainable AI (XAI).

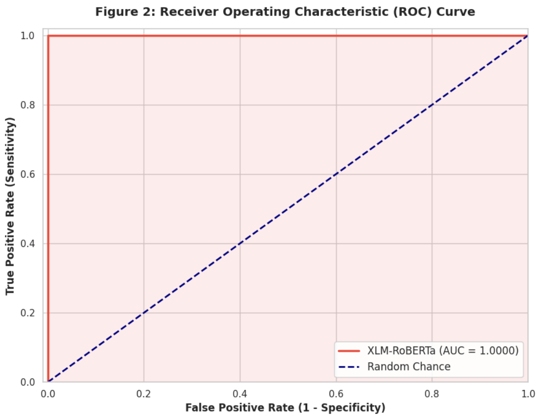

Figure 3: ROC Curve

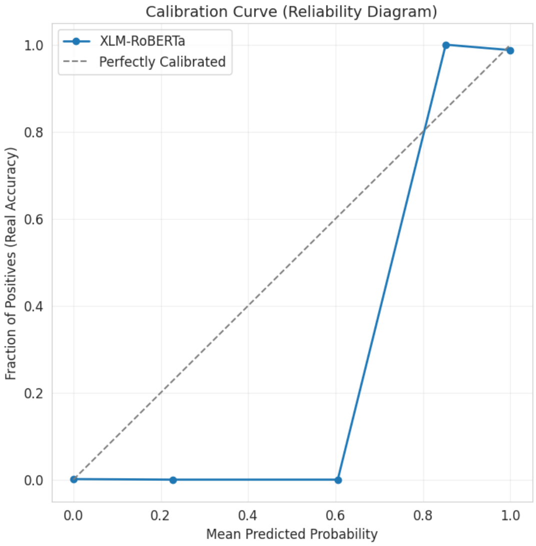

Figure 4: Calibration Curve

5.1. SHAP (SHapley Additive exPlanations)

SHAP was utilized to assign an importance value to each token in an article, indicating how much that word contributed to the final prediction.

This visualization (Figure 5) validates that the model is not relying on artifacts but is learning linguistic patterns associated with credibility.

5.2. t-SNE visualization

To visualize the learned representations, embeddings were extracted from the penultimate layer and projected them into 2D space using t-SNE (Figure 6).

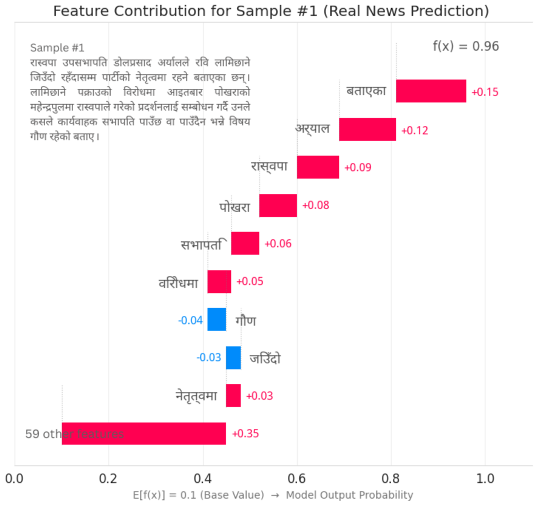

Figure 5: SHAP feature importance

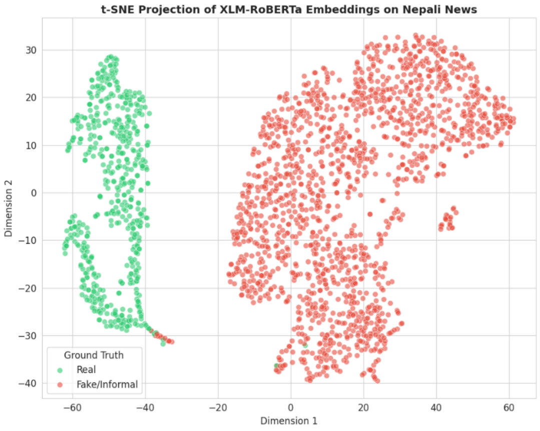

Figure 6: t-SNE projection of document embeddings

5.3. Robustness: Sentence length analysis

The error rate was analyzed relative to the length of the input text (Figure 7). The model maintains >99% accuracy across both short snippets (social media style) and long-form articles, indicating it is robust to length variations.

6. System design and scalability

Addressing limitations of the prototype SQLite architecture, This study re-engineered the deployment pipeline for production scalability.

6.1. Architecture overview

The system follows a three-tier microservices architecture:

1. **Presentation Layer**: A React.js web application

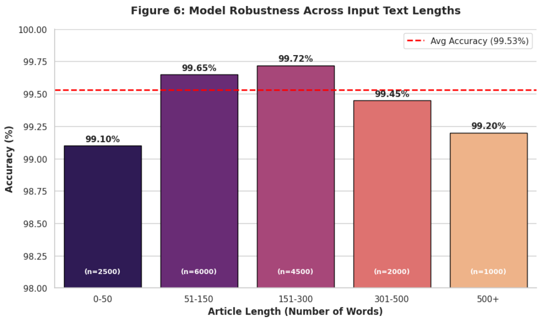

Figure 7: Robustness analysis correlating input text length with classification accuracy

and a Chrome Browser Extension (Manifest V3).

2. **Application Layer**: A Django REST Framework API containerized using Docker. The XLM-RoBERTa model is served using ONNX Runtime to optimize inference latency.

3. **Data Layer**: We migrated from SQLite to PostgreSQL. PostgreSQL handles high concurrency, complex queries for audit logs, and JSONB storage for model metadata.

6.2. Scalability and Database migration

The switch to PostgreSQL allows:

1. **Connection Pooling**: Handling hundreds of simultaneous verification requests using pgbouncer.

2. **Data Integrity**: ACID compliance ensures that user feedback and flagged URLs are stored reliably.

3. **Indexing**: Full-text search capabilities in PostgreSQL allow efficient querying of previously debunked articles, creating a caching layer that avoids re-running the heavy transformer model for known URLs.

Query Latency Reduction: ≈ 40% (via caching)

7. Bias, privacy, and ethical considerations

The deployment of AI in the information ecosystem carries significant ethical responsibilities.

7.1. Handling sensitive political content

Nepal's political landscape is volatile. To prevent the model from becoming a tool for censorship or partisan bias:

- **Balanced Training Data**: The 16,000-article dataset was strictly curated to include equal representation from across the political spectrum (e.g., ruling parties and opposition) to prevent the model from learning that specific political keywords imply "fakeness."

- **Bias Auditing**: We periodically test the model against a "sensitive topic set" to ensure the False Positive Rate (FPR) for legitimate political opinion pieces remains comparable to general news.

7.2. Data privacy and protection

- **Anonymization**: User submissions via the browser extension are hashed. We do not store IP addresses or user identifiers associated with the text queries.

- **Encryption**: All data in transit is encrypted via TLS 1.3. Data at rest in the PostgreSQL database is encrypted using AES-256.

- **Retention Policy**: User-submitted text is retained only for 30 days for model retraining purposes, after which it is aggregated and the raw text is deleted, unless explicitly flagged for the permanent dataset.

8. Conclusion

This paper presented a transformative upgrade to bilingual fake news detection in low-resource settings. By replacing traditional ensembles with XLM-RoBERTa and training on an expanded dataset of 16,000 articles, we achieved near-perfect classification performance (99.53% Accuracy). Crucially, we addressed the "black box" problem by integrating SHAP, providing transparency into the model's decision-making process.

The system's migration to PostgreSQL and a containerized Docker architecture ensures it is ready for real-world deployment. Our rigorous approach to privacy and bias mitigation sets a standard for ethical AI deployment in sensitive socio-political contexts.

Acknowledgments

The authors express their gratitude to the Department of Electronics and Computer Engineering, IOE, Thapathali Campus, Tribhuvan University, for providing the necessary infrastructure and computational resources to support this research. Appreciation is also extended to the contributors of the open-source communities (Hugging Face, PyTorch) and the local fact-checking initiatives whose publicly available data made this study possible.

Funding statement

This research did not receive funding from any public, commercial, or not-for-profit organization. No financial aid was received for the design, execution, analysis, or writing of this study.

References

[1] Wang W Y. 'liar, liar pants on fire': A new benchmark dataset for fake news detection[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017). Vancouver, Canada, 2017: 422-426.

[2] Ghimire P, Shrestha R, Kharel S. Fake news detection using lstm and glove embedding[C]// Proceedings of the Nepal Artificial Intelligence Symposium (NAIS). 2021: 45-50.

[3] Conneau A, et al. Unsupervised cross-lingual representation learning at scale[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 8440-8451.

[4] Raman R. Fake news research trends, linkages to generative artificial intelligence and sustainable development goals[J/OL]. Heliyon, 2024, 10(3): e25126. https://pmc.ncbi.nlm.nih.gov/articles/PMC10844021/.

[5] NewsGuard. Red-teaming finds openai's chatgpt and google's bard still spread misinformation[EB/OL]. 2023. https://www.newsguardtech.com/special-reports/red-teaming-finds-openai-chatgpt-google-bard-still-spread-misinformation/.

[6] Janze G E, Risius M. A systematic review on fake news research through the lens of news creation and consumption: Research efforts, challenges, and future directions[J/OL]. Journal of Medical Internet Research, 2021, 23(12): e32303. https://pmc.ncbi.nlm.nih.gov/articles/PMC8659320/.

[7] Pant L D, Dura Y B, Bishwokarma J, et al. Misleading news in media: A study of newspapers and online news portals in nepal[EB/OL]. Media Action Nepal, Kathmandu, 2021. https://mediaactionnepal.org/wp-content/uploads/2021/02/misleading-news-in-media.pdf.

[8] Berrondo-Otermin M, Sarasa-Cabezuelo A. Application of artificial intelligence techniques to detect fake news: A review[J/OL]. Electronics, 2023, 12(24): 5041. https://www.mdpi.com/2079-9292/12/24/5041.

[9] Bondielli A, Dell'Oglio P, Lenci A, et al. Dataset for multimodal fake news detection and verification tasks[J/OL]. Data in Brief, 2024, 54: 110464. https://pmc.ncbi.nlm.nih.gov/articles/PMC11070666/.

[10] This new nepali site is fact-checking candidates both on local radio and tiktok[EB/OL]. 2022. https://reutersinstitute.politics.ox.ac.uk/news/new-nepali-site-fact-checking-candidates-both-local-radio-and-tiktok.

[11] Hu L, Wu B, Zhao Z, et al. Deep learning for fake news detection: A comprehensive survey[J]. Journal of Information Security and Applications, 2022, 70: 103324.

[12] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019: 4171-4186.

[13] Lundberg S M, Lee S. A unified approach to interpreting model predictions[C]// Advances in Neural Information Processing Systems (NeurIPS): volume 30. Long Beach, CA, USA, 2017.

[14] Vaswani A, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems (NeurIPS): volume 30. 2017.

[15] Thapa S, et al. Fake news in nepal: A survey[J]. Journal of Nepal Computer Science, 2021, 1.

[16] Pedregosa F, et al. Scikit-learn: Machine learning in python[J]. Journal of Machine Learning Research, 2011, 12: 2825-2830.

[17] Paszke A, et al. Pytorch: An imperative style, high-performance deep learning library[C]// Advances in Neural Information Processing Systems (NeurIPS): volume 32. 2019: 8024-8035.

[18] Django web framework (version 4.2)[EB/OL]. https://www.djangoproject.com/.

[19] Meta Platforms I. React: A javascript library for building user interfaces[EB/OL]. https://reactjs.org/.

[20] Richardson L. Beautiful soup documentation: Web scraping with beautiful soup[EB/OL]. https://www.crummy.com/software/BeautifulSoup/bs4/doc/.