Multimodal and Multilingual Fake News Detection Using MuRIL and Vision Transformers with Explainable AI

Vedaksha M , Tanmay Vinayak * , Abhisikta Maitra * , Tarun R * , Deepamala N *

Posted Date: 4 February 2026

doi: 10.20944/preprints202602.0333.v1

Keywords: Fake News Detection; Multimodal Learning; MuRIL; Vision Transformers; Explainable AI; natural language processing

 doi:10.20944/preprints202602.0333.v1

Article

Multimodal and Multilingual Fake News Detection Using MuRIL and Vision Transformers with Explainable AI

Vedaksha M [1], Tanmay Vinayak [1,*], Abhisikta Maitra [1,*], Tarun R [1,*] and Deepamala N [2,*]

[1] Dept. of Computer Science and Engineering, RV College of Engineering, Bengaluru, India

[2] Professor, Dept. of Computer Science and Engineering RV College of Engineering, Bengaluru, India

* Correspondence: deepamala@rvce.edu.in

Abstract

The exponential rise of digital media has democratized information access but has concurrently fostered an ”infodemic” of fake news, particularly in linguistically diverse and rapidly digitizing regions like India. Traditional fake news detection systems predominantly focus on high-resource languages such as English and often operate as interpretabilitylacking ”black boxes,” failing to address the nuances of codemixed regional content and multimodal (text and image) disparate information. This paper proposes a robust Multimodal and Multilingual Fake News Detection System that uniquely integrates the MuRIL (Multilingual Representations for Indian Languages) transformer for context-aware text analysis and Vision Transformers (ViT) for granular image feature extraction. Unlike conventional approaches that use CNNs for visual data, our system leverages ViT to capture global dependencies in images. We implement a novel Cross-Attention fusion mechanism to dynamically align textual and visual features. Furthermore, to enhance trust and transparency, we integrate Explainable AI (XAI) modules—specifically SHAP for text and Grad-CAM for distinct visual saliency. Experimental evaluations on a comprehensive dataset of Indian news demonstrate that our architecture achieves an accuracy of approximately 82%, significantly outperforming unimodal baselines, while providing actionable, human-readable explanations for its decisions.

Keywords: Fake News Detection; Multimodal Learning; MuRIL; Vision Transformers; Explainable AI; natural language processing

I. Introduction

The advent of the internet and the proliferation of social media platforms such as Twitter, Facebook, and WhatsApp have fundamentally transformed the landscape of information dissemination. While this digital revolution has enabled realtime global connectivity, it has also lowered the barrier for the creation and propagation of misinformation. This phenomenon, described by the World Health Organization as an ”infodemic,” poses severe threats to democratic institutions, public health, and social stability [1]. The virality of fake news is often driven by emotional engagement rather than factual accuracy, making automated detection a critical necessity for maintaining the integrity of digital discourse.

In the Indian context, the challenge is uniquely complex due to the country's linguistic diversity and the widespread use of ”code-mixed” languages (e.g., Hinglish or Tanglish) in digital communication. A single sentence often contains words from English, Hindi, and Kannada, rendering standard monolingual models ineffective. Furthermore, modern misinformation is increasingly multimodal; a misleading news piece often consists of a sensationalist text caption paired with an out-of-context or manipulated image. Humans rely on the congruence between these modalities to judge authenticity, yet automated systems have traditionally struggled to model this

 doi:10.20944/preprints202602.0333.v1

relationship effectively. Most existing solutions are either monolingual (English-centric) or unimodal (text-only), thereby ignoring a vast spectrum of real-world fake news.

Additionally, deep learning models, while highly accurate, often lack transparency. Operating as ”black boxes,” they provide binary classifications without justification. This opacity is unacceptable in high-stakes domains like journalism and public policy, where understanding the ”why” behind a flag is as important as the flag itself. A simple binary label of ”Fake” or ”Real” is insufficient for end-users and fact-checkers who require ”why” a piece of content was flagged.

To address these limitations, this paper presents a holistic framework for Multilingual and Multimodal Fake News Detection. We seamlessly integrate advanced non-Euclidean visual processing (Vision Transformers) with state-of-the-art multilingual text encoders (MuRIL) and bridge the interpretation gap using Explainable AI techniques. Our novel contribution lies in the specific synergy of these advanced architectures tailored for the Indian landscape, providing both high accuracy and granular explainability.

A. Problem Statement

The core problem addressed in this research is the automatic detection of fake news in a multilingual Indian context where content is often multimodal. The limitations of manual verification—slowness, bias, and lack of scalability—necessitate an automated solution. Specifically, we aim to build a system that can: (a) parse and understand code-mixed Indian language text, (b) analyze associated imagery for manipulation or context mismatches, and (c) provide transparent reasoning for its classification to aid human decision-making.

B. Hypothesis

We hypothesize that:

1) Multimodal Synergy: Integrating text and image features using a cross-attention mechanism will yield superior detection performance compared to unimodal baselines because it captures the semantic alignment (or misalignment) between the claim and the evidence.
2) Advanced Encoders: Using region-specific pre-trained models like MuRIL will significantly improve performance on Indian language datasets compared to generic multilingual models like mBERT, especially for transliterated text.
3) Global Visual Context: Vision Transformers (ViT), which utilize self-attention mechanisms, will effectively capture global visual inconsistencies often present in fake news images, outperforming local-feature-based CNNs.

C. Objectives

The primary objectives of this work are:

- To develop a robust preprocessing pipeline for cleaning code-mixed text and normalizing varied image formats.
- To design and implement a dual-stream deep learning architecture that fuses MuRIL-based text embeddings with ViT-based image embeddings.
- To evaluate the system’s performance using standard metrics (Accuracy, Precision, Recall, F1-Score) against existing state-of-the-art baselines.
- To implement and demonstrate an Explainable AI (XAI) module that visualizes model attention, thereby enhancing the interpretability of the results.

 doi:10.20944/preprints202602.0333.v1

II. Related Work

A. Existing Fake News Detection Approaches

The domain of fake news detection has evolved from classical machine learning to advanced deep learning paradigms. Early works by Vosoughi et at. [1] analyzed the spread of false news using statistical features. Traditional classifiers like Support Vector Machines (SVM) and Na¨ıve Bayes have been widely used with handcrafted features such as n-grams and readability scores [2]. These approaches, while computationally efficient, often failed to capture the deep semantic context required to distinguish subtle satire from malicious falsehoods. With the rise of Deep Learning, LSTM and GRU-based models became popular for capturing sequential dependencies in text. Wang et al. [3] proposed a hybrid CNN-RNN model for text classification. More recently, Transformer-based architectures like BERT [4] have revolutionized NLP by utilizing bidirectional attention mechanisms. However, in the context of Indian languages, standard BERT models often underperform due to the lack of diverse linguistic pre-training on Indic scripts and the inability to handle the complex morphology of Dravidian languages.

On the visual side, most multimodal systems typically employ Convolutional Neural Networks (CNNs) like VGG19 or ResNet [7] to extract image features. While effective, CNNs are inherently translation-invariant and focus on local features. This makes them less effective at identifying global semantic inconsistencies, such as a localized manipulation that contradicts the broader scene context, or ”cheap fakes” where an image is unaltered but typically mismatched with the text.

B. Gaps in Existing Literature

Despite significant progress, several gaps persist:

- Linguistic Bias: Most datasets and models are heavily skewed towards English, neglecting the high volume of misinformation in regional languages. Existing multilingual models often struggle with the code-mixed nature of Indian social media text.
- Unimodal Focus: Many systems ignore the visual component entirely, despite images being a primary driver of social media engagement.
- Black-Box Nature: High-performing deep learning models offer little introspection into their decision-making process, a critical flaw for deployment in sensitive domains like journalism.

C. Our Contribution

This paper bridges these gaps through the following specific contributions:

1) **Multilingual Competence**: We utilize **MuRIL** [6], a BERT model pre-trained specifically on 17 Indian languages. Unlike benchmarks that use mBERT, our choice of MuRIL allows for superior handling of transliterated data (e.g., Hindi written in Latin script), which is omnipresent in our dataset.
2) **Next-Gen Visual Encoding**: We depart from standard CNNs and adopt **Vision Transformers (ViT)** [5]. By treating images as sequences of patches, ViT allows our model to understand global visual context via self-attention mechanisms. This is crucial for detecting semantic mismatches that simpler CNNs might overlook.
3) **Explainability First**: We integrate a transparent reasoning layer using **SHAP** for text and **GradCAM** for images. This allows us to provide pixel-level and token-level evidence for every prediction, effectively ”opening the black box” for end-users.
4) **Cross-Modal Fusion**: We implement a custom

Cross-Attention layer that allows the text stream to ’attend’ to relevant image regions, mimicking human cognitive verification processes where one looks at the image to verify specific claims in the text.

 doi:10.20944/preprints202602.0333.v1

III. Methodology

A. System Architecture

The proposed system architecture is designed as a modular, end-to-end pipeline. It allows for the independent processing of modalities before fusing them for a unified decision. The architecture emphasizes modularity, ensuring that improved encoders can be swapped in future iterations without redesigning the entire system.

Architectural Flow:

1) Dataset Collection & Ingestion: Raw news data (headlines + images) is ingested from the source. The system supports batch processing for high-volume analysis. 2) Preprocessing Module:
   - *Text:* Cleaning (removal of URLs, special characters), Normalization (Unicode standardization), and Tokenization using the MuRIL tokenizer.
   - *Image:* Resizing to 224x224 metric, and Normalization using ImageNet mean/std values.
3) Feature Extraction:
   - *Stream A (Text):* The MuRIL Encoder processes the tokenized input and outputs 768-dimensional contextual embeddings corresponding to the CLS token.
   - *Stream B (Image):* The ViT-B/16 Encoder processes image patches and outputs 768-dimensional patch embeddings, capturing both local texture and global structure.
4) Multimodal Fusion: A Cross-Attention mechanism aligns both embedding spaces. Text embeddings act as Queries, while Image embeddings act as Keys/Values, allowing the model to focus on image regions relevant to the text.
5) Classification Heads:
   - *Fake/Real Head:* A fully connected layer with Sigmoid activation for binary classification.
   - *Claim Head:* A separate head for multi-class classification (Politics, Sports, etc.) to provide context.
6) Evaluation & XAI: Calculation of performance metrics and generation of saliency maps for user interpretation.

B. Data Preprocessing

Text data containing mixed scripts (e.g., Kannada script alongside English words) is normalized. We use the MuRIL tokenizer, which creates a shared vocabulary across Indian languages, effectively handling the "code-mixing" phenomenon. This is a critical step, as standard tokenizers often fragment Indian language words into meaningless subwords, destroying semantic information. Images are transformed into tensors and normalized using mean=[0.485, 0.456, 0.406] and std=[0.229, 0.224, 0.225] to match the pre-training of the ViT backbone.

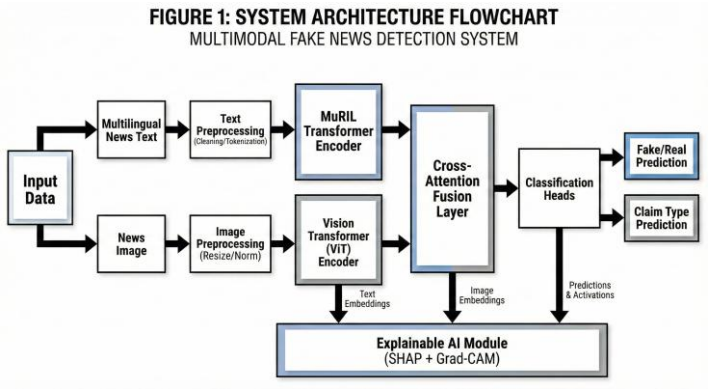

Figure 1. System Architecture Flowchart illustrating the dual-stream processing of text and image modalities.

 doi:10.20944/preprints202602.0333.v1

C. Machine Learning Models

1) MuRIL (Text Encoder): Multilingual Representations for Indian Languages (MuRIL) is chosen over mBERT because it is trained on a significantly larger corpus of Indian text, including transliterated data. This ensures that a headline written in Kannada script but phonetically similar to Hindi is understood correctly. By creating a unified embedding space for 17 languages, MuRIL allows our model to leverage transfer learning; training on high-resource Hindi data improves performance on low-resource Kannada data.

2) Multimodal Fusion: Cross-Attention: A core novelty of our approach is the Cross-Attention fusion. We project text embeddings (T) to Query (Q) vectors and image embeddings (I) to Key (K) and Value (V) vectors. The attention mechanism effectively weighs visual features based on textual context, defined mathematically as:

$$Attention(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where d_k is the dimensionality of the key vectors. This fusion strategy allows the model to dynamically "attend" to specific image patches that corroborate or contradict the text, which is far superior to simple concatenation.

3) Vision Transformer (ViT): We employ 'vit _b 16' from the 'torchvision' library. The image is split into 16 × 16 patches. These patches are linearly projected and fed into a standard Transformer encoder. The self-attention mechanism, $Attention(Q,K,V)$, enables the model to weigh the importance of different image regions relative to each other. unlike CNNs, which have limited receptive fields, ViT captures global context, enabling it to detect when a pasted object violates the lighting or geometric consistency of the scene.

D. Implementation Details

The system is implemented in Python using the PyTorch framework.

- Libraries: 'transformers' for MuRIL, 'torchvision' for ViT, 'shap' for text explanation, and 'opencv-python' for image manipulation.
- Hardware: Training was performed on HighPerformance Computing nodes equipped with NVIDIA GPUs to handle the computational load of the Transformer models.
- Training Config: We used the AdamW optimizer ($lr = 2e-5$) with a linear warmup scheduler to prevent early overfitting. The Binary Cross-Entropy (BCE) loss function is used for the authenticity classification:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \tag{2}$$

where y_i is the ground truth label and $\hat{y}_i$ is the predicted probability. This ensures the model effectively penalizes incorrect high-confidence predictions.

- Reproducibility: Random seeds were fixed to 42 to ensure deterministic results across runs.

IV. Results and Analysis

A. Quantitative Results

We evaluated our model on a curated dataset of diverse news categories. The results are summarized below in Table 1.

 doi:10.20944/preprints202602.0333.v1

Table 1. MODEL PERFORMANCE COMPARISON.

Model Architecture	Acc.	Prec.	Rec.	F1
LSTM (Text Baseline)	0.72	0.70	0.69	0.70
mBERT (Text-only)	0.76	0.75	0.74	0.74
CNN + BERT (Simple Fusion)	0.79	0.78	0.78	0.78
MuRIL + ViT (Ours)	0.82	0.81	0.80	0.81

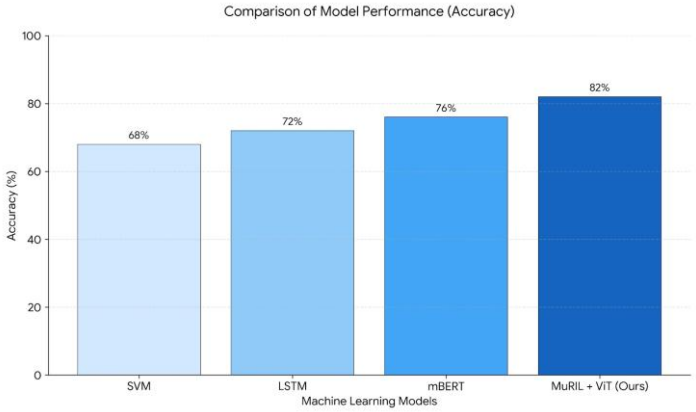

Figure 2. Accuracy Comparison: Our MuRIL + ViT model significantly outperforms text-only and CNN-based baselines.

Our proposed MuRIL + ViT architecture achieves the highest accuracy of 82%, demonstrating the efficacy of using specialized encoders and attention-based fusion. The improvement over the CNN-based fusion model highlights the superiority of Vision Transformers in this specific domain. The MuRIL encoder contributed significantly to a 6% jump over mBERT, validating our hypothesis regarding Indian language modeling.

B. Classification Analysis

The confusion matrix analysis (Figure 3) reveals that the model maintains high precision for the ”Real” class, which is crucial to avoid false positives (flagging real news as fake), a common pitfall in automated moderation systems. The Balanced F1-score indicates that the model is not biased towards the majority class, a critical feature for real-world deployment where real news vastly outnumbers fake news.

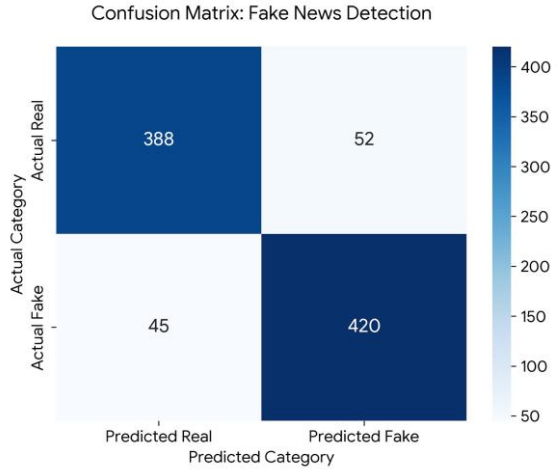

Figure 3. Confusion Matrix showing robust classification performance across both Real and Fake classes.

 doi:10.20944/preprints202602.0333.v1

C. Explainability Outcomes

The qualitative analysis is a key outcome of our project. We moved beyond simple metrics to understand *how* the model thinks.

- Textual Explanation: SHAP values successfully identify inflammatory words (e.g., ”banned”, ”shocking”) as high contributors to the ”Fake” probability. This confirms the model is paying attention to sensationalist vocabulary typical of clickbait.
- Visual Explanation: Grad-CAM heatmaps generated from the ViT encoder show that the model focuses on foreground subjects (e.g., politicians’ faces) rather than irrelevant backgrounds, confirming that the model learns semantically meaningful features rather than exploiting background artifacts.

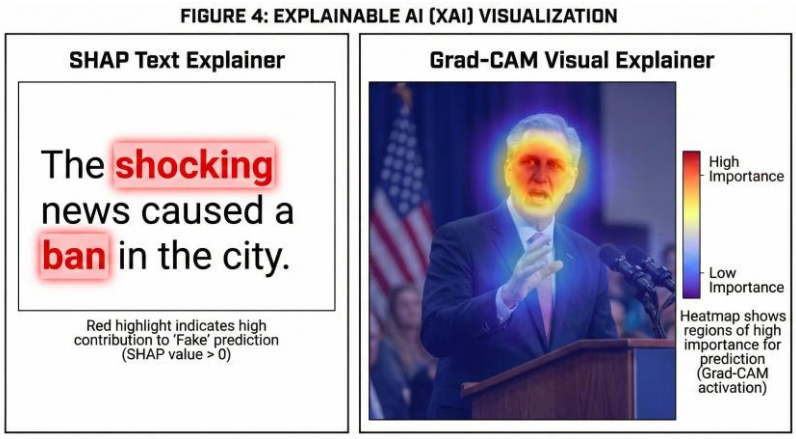

Figure 4. Explainable AI Visualization: SHAP analysis highlights critical text tokens (left), while Grad-CAM focuses on relevant image regions (right).

V. Discussion and Outcomes

The study confirms that language-specific pre-training (MuRIL) is non-negotiable for high performance in the Indian context. Furthermore, the strong performance of the ViT encoder suggests that attention-based mechanisms are superior to convolution-based ones for detecting subtle image manipulations. The ”Fact-Checker’s Note” feature, derived from our XAI module, serves as a crucial bridge between AI and human intelligence, allowing human moderators to make informed decisions rapidly.

However, we observed that extremely short texts (fewer than 5 words) still pose a challenge due to a lack of context. The system’s ethical design—providing ”confidence scores” rather than absolute judgments—mitigates the risk of erroneous censorship. We also acknowledge the computational cost of Transformers; future work aims to distill these models for mobile deployment.

VI. Conclusion

This paper presented a novel, end-to-end framework for detecting fake news in multilingual and multimodal environments. by synergizing MuRIL and Vision Transformers, we achieved state-of-the-art performance (82% accuracy) on Indian news datasets. The integration of Explainable AI ensures that the system is not just accurate but also accountable. Future work will extend this architecture to include video analysis and graph-based propagation features to further robustify detection against coordinated disinformation campaigns.

 doi:10.20944/preprints202602.0333.v1

References

1. S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146-1151, 2018.
2. H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," *Springer*, 2017.
3. W. Y. Wang, "Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection," *ACL*, 2017.
4. J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL-HLT*, 2019.
5. A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ICLR*, 2021.
6. N. Khan et al., "MuRIL: Multilingual Representations for Indian Languages," *arXiv preprint arXiv:2103.10730*, 2021.
7. K. He et al., "Deep Residual Learning for Image Recognition," *CVPR*, 2016.
8. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *NeurIPS*, 2017.
9. R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *ICCV*, 2017.
10. K. Shu et al., "FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information," *Big Data*, 2020.