Multimodal Fake News Detection in Bilingual Media: A Vision–Language Fusion Approach Using Task-Specific CNNs and Multilingual Transformers

Anonymous ACL submission

Abstract

Misinformation on social media increasingly exploits visually persuasive thumbnail-style content, yet multimodal fake news detection remains underexplored for low-resource languages. This paper presents a bilingual multimodal framework for fake news detection using textual and visual information extracted from Bangla and English social media thumbnails collected from YouTube, Facebook, and Instagram. A manually annotated dataset of 19,890 thumbnail images with embedded textual content is constructed. Multilingual transformer models are fine-tuned using a robust text preprocessing pipeline, while a custom convolutional neural network is designed for visual feature extraction. Unimodal predictions are combined using a late fusion strategy to enhance robustness. Experimental results demonstrate that the XLM-RoBERTa-base model outperforms other multilingual transformer-based models on OCR-extracted text, achieving the highest macro F1-score of 97.20. The Visual Misinformation Detection Convolutional Neural Network (VMD-CNN) model achieves 96.05% accuracy on visual content, confirming that textual cues dominate while visual features provide complementary signals. Late fusion of the best-performing unimodal models further improves overall performance, reaching 97.15% test accuracy, highlighting the effectiveness of decision-level fusion in integrating heterogeneous modalities. This framework provides a practical solution for detecting misleading social media content, particularly in bilingual low-resource settings, and offers a foundation for future research in multilingual multimodal fake news detection and automated content moderation.

1 Introduction

Fake news detection has become drastically more important in this digitally evolved era due to the rapid spread of information through social media platforms, online news channels, and instant messaging networks, where false and misleading content can influence public opinion, social harmony, and decision-making mechanisms instantaneously. The massive volume and rate of spreading digital content make it nearly impossible to detect their authenticity, increasing the risk of misinformation affecting political, health, and economic information. Misinformation is often difficult to detect because it is shared through Facebook or Instagram posts, YouTube channel thumbnails, memes, or combinations of text and images. Artificial intelligence plays a vital role in addressing this challenge by enabling automated analysis of large-scale data using machine learning, deep learning, and natural language processing techniques to identify patterns, linguistics, and sources that diffuse fake news, supporting faster, more accurate, and scalable misinformation detection systems. In a country like Bangladesh, where political situations are unpredictable and chaotic, fake news can make the situation even worse. As a solution to these conundrums, our proposed model for a fake news detection system can contribute significantly. Our model uses both English and Bangla textual and visual input for classification and uses the outputs for late fusion, which gives the final verdict.

Previously, the problem has been addressed and attempted to be solved using machine learning techniques (Capuano, 2023), natural language processing models (Reddy, 2024), and also a hybrid of multiple models for the architectural advantages of specific models (Segura-Bedmar and Alonso-Bartolome, 2021). Although the dataset used and the approach of detection differed from unimodal approaches to multimodal approaches. For both image and text datasets, researchers have seen excellent results. For unimodal approaches, the Bangla language dataset has been used, but image-based datasets are one of the least available datasets; the ones that exist are mostly from news articles and

fake images only, whereas our dataset contains YouTube thumbnails for both Bangla and English languages. In this age of social media, thumbnails play the main role in getting views, yet this area remains unexplored for both the English and Bangla languages.

The main contributions of this work are as follows:

- Introduced bilingual multimodal fake news detection using social media thumbnails (YouTube, Facebook, Instagram) in Bangla/English, addressing the scarcity of Bangla image-based work
- Created a comprehensive dataset of 19,890 thumbnails and used OCR for combined visual-textual decisions.
- Built multilingual text preprocessing and VMD-CNN, outperforming pretrained models on thumbnail visuals.
- Applied late fusion of unimodal predictions, handling weak image-text alignment realistically.
- Achieved 97.15% accuracy, surpassing unimodal baselines through complementary visual/textual cues.

2 Related Work

Detection of fake news has always been a hot topic in regard to internet news, making the research field enthusiastic to work on it. Researchers used both textual and visual modalities for detecting misleading content. Before social media platforms supported uploading images or videos, texts were the main content on such platforms, contributing to a huge dataset to train the models. Advancements in social media allowed multimedia uploads, giving multimedia datasets. In recent times, deep learning models have significantly dominated computer vision and NLP, which further assists fake news detection. In this section, we review related works by examining their modalities, methodologies, and linguistic aspects.

2.1 Text-Based Fake News Detection

Often, studies focus only on textual data and traditional machine learning models like Naive Bayes, SVM, and Logistic Regression on datasets similar to Kaggle Fake News and LIAR. Deep learning models have been proven to be superior in performance for text-based detections. Saleh et. al.(Saleh et al., 2021) in 2021 put forward OPCNN-FAKE, an optimized CNN for text classification. Kaliyar et. al.(Kaliyar, 2020) in 2020 proposed FNDNet, a deep CNN using GloVe embeddings. Hybrid architectures similar to CNN-RNN and CNN-LSTM are also implemented to understand both infamous patterns and sequential dependencies (Nasir, 2020; Umer, 2020). All of these studies presented impressive results but are constrained by text-only representations and limited lingual diversity.

2.2 Multimodal Fake News Detection

There are multiple studies that implemented the fusion of visual and textual datasets to develop a more efficient fake news detection system. Segura-Bedmar et. al. (Segura-Bedmar and Alonso-Bartolome, 2021) proposed a bilingual multimodal system using the models BiLSTM+CNN and BERT on the Fakeddit dataset, limited to the English language but not available for Bangla fake news. In 2020, Giachanou et. al. (Giachanou et al., 2020) proposed a multimodal approach that separates images and texts from newspaper articles and that uses BERT for text and VGG16 with LSTM for images and then fuses them via an attention mechanism. Following Giachanou's approach in 2022, Jing et. al. (Jing et al., 2022) implemented a progressive fusion that combined BERT-based text features and Swin Transformer-based visual features. The approach of Sharma et. al. (Singh, 2021) in 2021 uses EfficientNetB0 for image features and RoBERTa for text, fused by late fusion, and adds Error Level Analysis (ELA) to detect fabrication in images. Zhang et. al.(Zhang, 2023) in 2023 modeled an image-text similarity using ResNet101, SBERT, and ViLBERT to check for inconsistencies between textual and visual data. Limitations still exist with respect to the Bangla language.

2.3 Other Relevant Works

Bilingual fake news detection is limited, mostly relying on TF-IDF and n-grams with ML models for Bengali and English, excluding visual data (Capuano, 2023). LLM-based approaches can augment text for low-resource languages like Bangla but remain unimodal (Bourgonje et al., 2017; Shibu, 2025). Multimodal methods, explored in hate speech detection with models like MMBT, ViLBERT, and Visual BERT, show promise but are mainly English-focused, and deep learning mod-

els are constrained by limited multilingual datasets and explainability (Bradley, 1997; Reddy, 2024).

3 Dataset Development

3.1 Data Collection

We manually collected 19,890 social media thumbnails from YouTube, Facebook, and Instagram, sourced from verified real news outlets and known misinformation-focused channels in both Bangla and English. The English subset comprises thumbnails from established international and regional news sources as well as English-language channels identified as persistent misinformation publishers. Similarly, the Bangla subset includes thumbnails from reputable Bangladeshi news and television channels, alongside Bangla-language sources known to disseminate misleading or fabricated content. A complete list of source channels for each language and class is provided in the Appendix.

Video identifiers were retrieved using Python scripts and subsequently used to automatically download the corresponding thumbnails. To reflect real-world misinformation dynamics, the dataset intentionally includes weakly aligned or semantically inconsistent images with embedded textual content—a common characteristic of online fake news where visual and textual cues are not strictly correlated. Figures 1 and 2 show representative samples from our collected dataset, including both fake and real news examples, and Figure 3 shows the extracted text samples. Table 1 represents the source channels used for thumbnail collection.

(a) English real news

(b) English fake news

Figure 1: Text-embedded thumbnails from the English subset.

(a) Bangla real news

(b) Bangla fake news

Figure 2: Text-embedded thumbnails from the Bangla subset.

3.2 Data Preprocessing

3.2.1 Textual Data Preprocessing

Text embedded in thumbnails was extracted using Google Gemini OCR and stored in a structured CSV. Preprocessing steps included URL removal, emoji and symbol filtering, punctuation normalization, lowercasing, and whitespace correction. Duplicate records were removed to prevent data leakage. Figure 4 shows the data distribution, where for the textual pipeline, the final corpus consists of 9,954 samples, comprising 5,419 Bangla and 4,535 English entries, with 5,447 labeled as fake and 4,507 as real. Pretrained multilingual tokenizers were employed for text processing, with a maximum sequence length of 512 tokens.

Language	Class	Example Text
Bangla	Fake	টাইম BANGLADESH ডঃ ইউনূস নির্দেশ পক্ষপাতব দলে ফিরবে সাকিব, তামিম, মাশরাফি ?
	Real	সেতু নির্মাণে ত্রুটি বেরিয়ে পড়েছে গার্ডারের রড — Maasranga HD
English	Fake	DH Russian jets to Iran leaked!
	Real	N.K.H.A Al Jazeera Afghanistan--Pakistan truce offers hope to stranded truckers

Figure 3: Extracted Text Sample

3.2.2 Image Data Preprocessing

For the image processing pipeline, we worked with a total of 9,936 images, split evenly between 4,968 fake and 4,968 real samples, having 2,484 images in Bangla and 2,484 in English, as shown in Figure 4. We resized all the images to 256×256 pixels and normalized the pixel values. To enhance efficiency, we utilized TensorFlow data pipelines with caching and prefetching features.

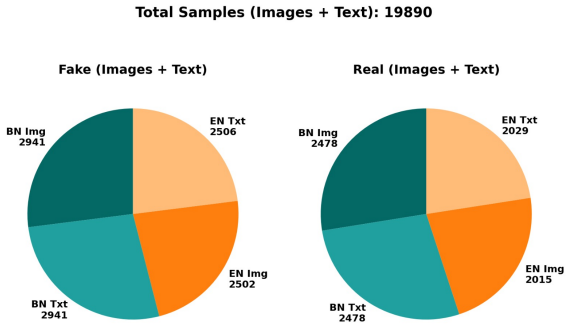

Figure 4: Data Distribution

3.2.3 Data Splitting

For both textual and visual pipelines, the datasets were independently partitioned using stratified

Language	Class	Source Channels
English	Real	5Pillars, AlJazeera, BangladeshPost, IslamChannel, MiddleEastEye, NBC-News, NewAgeBD, OwenJones, SecularTalk, TRTWorld, TheEconomist, The-FrontLines
English	Fake	americannews_en_fake, business_basic_en_fake, cipher_signal_en_fake, crime_watchers_en_fake, danny_haiphong_en_fake, mega_operations_en_fake, safety_zone_en_fake, the_left_lens_en_fake, truth_stream_en_fake, ukraine_today_en_fake
Bangla	Real	BangladeshProtidin, BusinessInspectionBD, BusinessStandard, DeshTV, EkusheyTV, JamunaTV, MaasrangaTV, SorwarAlam
Bangla	Fake	news_alert_bn_fake, newstoday_bn_fake, public360_bn_fake, pub-lic_insight_bn_fake, somoy_bn_fake, time24_bn_fake

Table 1: Source Channels Used for Thumbnail Collection

class-wise sampling to preserve balanced label distributions. Each modality was split into training, validation, and test sets in a 70:15:15 ratio, ensuring consistent class representation across all experimental phases and reducing potential training and evaluation bias.

4 Methodology

4.1 Model Overview

We explored a late fusion model that combines predictions from our VMD-CNN for images and XLM-RoBERTa for text to see if blending these two modalities improves performance. We paired the data by aligning the indices of image and text samples. We used the size of the smaller dataset to ensure that everything matched up properly, which helped us maintain reliable supervision for each fusion instance. This decision-level approach allows us to merge the probability outputs without needing to force the two feature sets—convolutional and transformer representations—to align perfectly. It helps us avoid conflicts while also enabling the model to determine the best weights for each modality through gradient-based optimization, even when the visual and textual elements don't perfectly correspond.

4.2 Unimodal Model Architecture and Training

4.2.1 Text-Based Classification Models

We utilized four multilingual transformer models—XLM-RoBERTa-base (XLM-R), BERT-base-multilingual-cased (m-bert-c), BERT-base-multilingual-uncased (m-bert-unc), and DistilBERT-base-multilingual-cased—as text encoders for a binary sequence classification task. Each model was initialized with pretrained weights and adapted by adding a classification layer with softmax activation to get class probabilities. Using the Hugging Face Models library (v4.x) for fine-tuning, we monitored performance through validation accuracy, precision, recall, and the macro-averaged F1-score. All models were subsequently evaluated on the test set for comparison. Our results showed that XLM-R outperformed the others, achieving a precision of 97.09%, which is better than other models. This positions XLM-R as a promising candidate for multimodal late fusion.

4.2.2 Image-Based Classification Models

For visual classification, five CNN variants—VMD-CNN (progressive filters 32→64→128 with max-pooling and flattening), AvgPool CNN (average pooling), GAP CNN (global average pooling), Constant Filters CNN (uniform 32–32–32 filters), and a Shallow CNN (two convolutional blocks)—were first evaluated. Among these, the Visual Misinformation Detection Convolutional Neural Network (VMD-CNN) achieved the highest performance and was therefore selected as the visual branch of the framework. VMD-CNN figure 6 is designed for thumbnail-level fake news classification and consists of three hierarchical convolutional blocks with progressively increasing feature depths (32→64→128), each integrating convolution, batch normalization, and max-pooling, followed by a dropout-regularized fully connected layer to produce compact visual embeddings for multimodal fusion. The selected VMD-CNN was further compared against pretrained architectures, including MobileNetV2, InceptionV3, Xception,

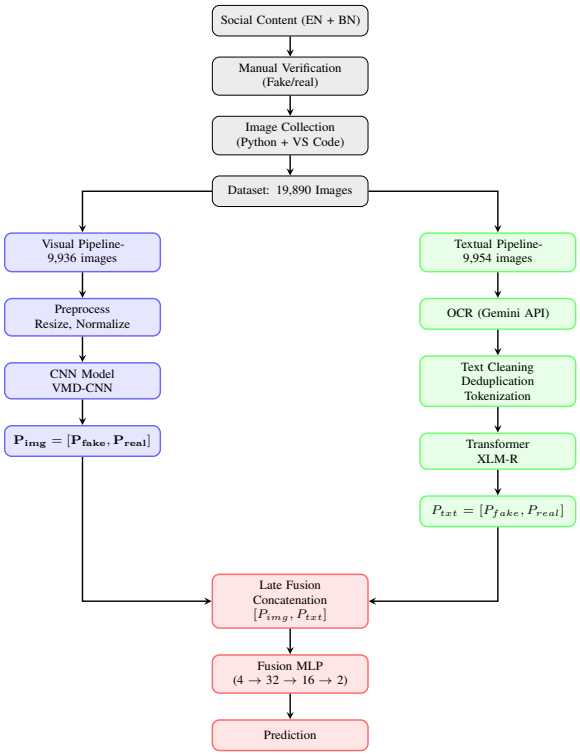

Figure 5: End-to-end workflow of the proposed multimodal fake news detection framework using late fusion.

and ResNet50, all initialized with ImageNet weights and used as frozen feature extractors with global average pooling. Trained using binary cross-entropy loss and the Adam optimizer with early stopping, VMD-CNN outperformed all pretrained models, achieving 96.05% accuracy, demonstrating the effectiveness of task-specific visual representations over generic object-centric features for fake news thumbnail classification.

4.3 Late Fusion Strategy

In our approach, we used decision-level fusion to combine predictions from different sources without forcing semantic alignment. Specifically, we took the probability outputs from both the image and text classifiers and concatenated them into a four-dimensional vector. This vector was then fed into a simple multilayer perceptron (MLP) designed with architecture $4 \to 32 \to 16 \to 2$.

$$\mathbf{f} = [p_{img}^{(0)}, p_{img}^{(1)}, p_{text}^{(0)}, p_{text}^{(1)}]$$

$$N_{fusion} = \min(N_{image}, N_{text})$$

To improve efficiency and reduce overfitting, only the fusion classifier was trained while the unimodal image and text models were kept frozen, enhancing robustness under weak visual–textual alignment.

4.4 Analysis and Performance Attribution

The proposed methodology emphasizes modularity and robustness for real-world bilingual fake news detection, where visual-textual alignment is often weak. Independent unimodal pipelines allow specialized optimization—transformers for linguistic nuances in OCR-extracted text and CNNs for visual artifacts—avoiding premature fusion that could propagate errors from noisy modalities.

4.4.1 Late Fusion Advantages

Decision-level fusion via concatenated probabilities and lightweight MLP (4→32→16→2) bypasses feature-space mismatches between convolutional and transformer architectures, enables modality weighting via backpropagation, and maintains interpretability. Freezing unimodal backbones prevents catastrophic forgetting.

4.4.2 Data and Preprocessing Rationale

Gemini OCR reliably handles thumbnail text; stratified 70-15-15 splits balance Bangla/English and fake/real classes. VMD-CNN targets thumbnail-specific cues (overlay patterns, compression) over ImageNet features, yielding 5%+ gains in ablations.

4.4.3 Performance Attribution

Textual cues predominantly drive detection (XLM-R: 97.09%), highlighting the role of linguistic irregularities, semantic discrepancies, and sensationalism in fake narratives. Visual evidence (VMD-CNN: 96.05%) offers supportive insights, even when relying on generic stock images. The fusion process (97.15%) combines these independent signals to enhance overall understanding, reducing the gap between training and testing accuracy (0.78%).

4.4.4 Summary of Key Insights

Overall, the integrated analysis yields three central findings:

- Task-specific architectural design is critical: VMD-CNN outperforms pretrained visual models in domain-specific fake news detection.
- Textual information: It is the primary discriminative modality, particularly for OCR-extracted bilingual data.

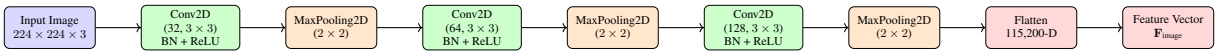

Figure 6: VMD-CNN architecture.

- Decision-level multimodal fusion: improves robustness and generalization without requiring strong visual-textual alignment.

These results validate the proposed methodology and highlight the effectiveness of late fusion for multilingual, low-resource, and weakly aligned multimodal fake news detection. Our codes are publicly available in this repository [1].

5 Results and Discussion

5.1 Image-Based Classification Performance

Table 2 offers a weighted performance overview of different convolutional neural network (CNN) architectures that were tested for detecting visual fake news. Four pretrained models—MobileNetV2, Xception, InceptionV3, and ResNet50—were benchmarked against the VMD-CNN optimized specifically for this task.

Table 2: Performances of CNN Models on Test Set

Models	Accuracy	Precision	Recall	Macro-F1
VMD-CNN	96.05	96.05	96.05	96.02
MobileNetV2	90.39	90.42	90.39	90.33
Xception	89.47	89.53	89.47	89.41
InceptionV3	88.16	88.29	88.16	88.11
ResNet50	82.96	83.25	82.96	82.68

The VMD-CNN surpassed the highest-ranked pretrained models, such as MobileNetV2 (90.39% accuracy), and achieved a 5.06% higher test accuracy of 96.05%. This gap raises the issue of whether ImageNet-pretrained features focus on objects rather than fake news cues like image artifacts, compression marks, text overlays, or odd compositions. It performed better than deep learning models in our experiments. These findings highlight the need for task-specific architectures in domain-focused tasks where generic features fall short. While prior studies (Kiela et al., 2020; Song et al., 2020) accomplished strong results using pretrained models, their dataset contained social media memes only in the English language, whereas our results are applicable for both the English and Bangla languages.

[1] https://github.com/seekersg96-png/Anonymous_Repo

5.2 Text-Based Classification Performance

Table 3 summarizes the weighted results of four multilingual transformer architectures after being fine-tuned for bilingual fake news detection. Each model was trained on carefully prepared textual data with balanced class distributions, and their effectiveness was evaluated on separate test sets that maintained natural class distributions to ensure realistic performance assessment.

Table 3: Performances of Multilingual Transformer Models

Model	Accuracy	Precision	Recall	Macro-F1
XLM-R	97.09	97.26	97.22	97.20
m-bert-unc	96.95	96.84	97.08	96.93
m-bert-c	96.82	96.87	96.82	96.80
DistilBERT	96.54	96.59	96.54	96.52

The XLM-R model stood out, achieving the highest performance across all evaluated metrics, with an accuracy of 97.09% and a macro F1-score of 97.20. This success can be attributed to its enhanced pretraining corpus and the Sentence Piece tokenization strategy, which significantly boosts its understanding of multilingual text. Although small, the improvement is substantial given high baseline performances compared to m-bert-unc (96.95%), m-bert-c (96.82%), and DistilBERT (96.54%). It is worth mentioning that all transformer-based text models achieve significantly higher performance than the best image-only model (VMD-CNN: 96.05%), indicating that textual features provide more discriminative information in the given dataset for fake news detection. This ranking implies the visual content encodes evidence in complement with textual narratives, such as semantic coherence, writing style, and factual consistency, to indicate content veracity. Previous BERT-based studies show promising results (Verma, 2022; Munir, 2022; Hossain et al., 2022), where our text data was extracted from YouTube thumbnails using Gemini OCR, illustrating robustness on semi-structured input.

5.3 Multimodal Late Fusion Performance

We explored a late fusion model that combines predictions from our VMD-CNN for images and XLM-R for text to see if blending these two modal-

Table 4: Performance of the multimodal fusion model.

Configuration	Val.	Test	Macro-F1
VMD-CNN + XLM-R	96.37	97.15	97.14

Val./Test = accuracy (%).

ities improves performance (Yu et al., 2025). As we detailed in Section III-C, we paired the data by aligning the indices of image and text samples. We used the size of the smaller dataset to ensure that everything matched up properly, which helped us maintain reliable supervision for each fusion instance. This decision-level approach allows us to merge the probability outputs without needing to force the two feature sets—convolutional and transformer representations—to align perfectly. It helps us avoid conflicts while also enabling the model to determine the best weights for each modality through gradient-based optimization, even when the visual and textual elements don't perfectly correspond.

Table 4 shows the results of the optimized fusion model, achieving a test accuracy of 97.15% with a 97.14 macro F1-score, recording a small but consistent increase over the best standalone text model (XLM-R: 97.09%) and a wider gap in comparison to the best image model (VMD-CNN: 96.05%). The results demonstrate that decision-level late fusion effectively integrates heterogeneous modalities, where textual features dominate prediction performance and visual features contribute complementary evidence across languages.

Conclusions

This paper introduced a bilingual multimodal fake news detection framework that combines OCR-extracted textual content and visual information from Bangla and English social media thumbnails. The proposed approach addresses the limited availability of image-based fake news studies for low-resource languages and is designed to operate under weak visual–textual alignment common in real-world social media content. Experimental evaluation demonstrates that textual features are the primary source of discriminative power, with XLM-R achieving a macro F1-score of 97.20, while the task-specific VMD-CNN outperforms ImageNet-pretrained models on thumbnail images. Decision-level late fusion of unimodal predictions further improves robustness, achieving a test accuracy of 97.15%. Overall, the framework provides a practical and scalable baseline for bilingual multimodal misinformation detection and contributes both a curated dataset and validated methodology for future research in multilingual automated content moderation. Overall, the framework provides a practical and scalable baseline for bilingual multimodal misinformation detection and contributes both a curated dataset and validated methodology for future research in multilingual automated content moderation.

Limitations

Despite its strong performance, the proposed framework has certain limitations. The model focuses on decision-level fusion and does not explicitly model fine-grained cross-modal interactions between textual and visual features. In addition, the analysis is restricted to static social media thumbnails and OCR-extracted text, without incorporating temporal video information, user engagement signals, or platform-specific propagation patterns. The evaluation is also limited to a bilingual Bangla–English setting. Future works can explore cross-attention or contrastive multimodal alignment techniques on fully aligned datasets to capture deeper semantic relationships between images and text. Extending the framework to video-level misinformation detection and incorporating contextual signals such as comments, sharing behavior, or metadata could further improve robustness. Expanding the dataset to include additional low-resource languages and experimenting with hybrid or ensemble multimodal architectures would enhance the system's global applicability and practical impact in automated misinformation monitoring and content moderation systems.

References

Peter Bourgonje, J. Moreno Schneider, and Georg Rehm. 2017. From clickbait to fake news detection: An approach based on detecting the stance of headlines to articles. In *Proceedings of the EMNLP Workshop on Natural Language Processing Meets Journalism*, pages 84–89.

Andrew P. Bradley. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

Nicola Capuano. 2023. Content-based fake news detection with machine and deep learning: A systematic review. *Neurocomputing*.

Anastasia Giachanou, Guobiao Zhang, and Paolo Rosso. 2020. Multimodal multi-image fake news detection.

In *Proceedings of the IEEE International Conference on Data Mining Workshops*.

Md Zakir Hossain, Ghulam Muhammad, Md Rafiul Islam, and Mamoun Alazab. 2022. CB-Fake: A multimodal deep learning framework for automatic fake news detection using capsule neural networks and BERT. *Multimedia Tools and Applications*, 81(5):5587–5620.

Jing Jing, Hongchen Wu, Jie Sun, Xiaochang Fang, and Huaxiang Zhang. 2022. Multimodal fake news detection via progressive fusion networks. *Information Processing & Management*.

Rohit Kumar Kaliyar. 2020. FNDNet: A deep convolutional neural network for fake news detection. *Computer Communications*.

Douwe Kiela and 1 others. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*.

Saad Munir. 2022. BiL-FaND: Leveraging ensemble techniques for efficient bilingual fake news detection. *International Journal of Machine Learning and Cybernetics*.

Jamal Abdul Nasir. 2020. Fake news detection: A hybrid CNN-RNN based deep learning approach. *Information Processing & Management*.

Suresh Reddy. 2024. Fake news detection using machine learning and NLP. *arXiv*.

Hager Saleh and 1 others. 2021. OPCNN-FAKE: Optimized convolutional neural network for fake news detection. *IEEE Access*.

Isabel Segura-Bedmar and Santiago Alonso-Bartolome. 2021. Multimodal fake news detection. *arXiv*, abs/2112.04831.

Hrithik Majumdar Shibu. 2025. From scarcity to capability: Empowering fake news detection in low-resource languages with LLMs. *arXiv*, abs/2501.09604.

Bhuvanesh Singh. 2021. Predicting image credibility in fake news over social media using a multimodal approach. *Neural Computing and Applications*.

Chenguang Song, Nianwen Ning, Yunlei Zhang, and Bin Wu. 2020. A multimodal fake news detection model based on cross-modal attention residual and multichannel convolutional neural networks. *Computer Networks*, 182:107476.

Muhammad Umer. 2020. Fake news stance detection using deep learning architecture (CNN-LSTM). In *Proceedings of an IEEE International Conference*.

Pawan Kumar Verma. 2022. Multi-modal message credibility for fake news detection using BERT and CNN. *Journal of Ambient Intelligence and Humanized Computing*.

Kai Yu, Shiming Jiao, and Zhilong Ma. 2025. Fake news detection based on BERT multi-domain and multimodal fusion networks. *Computer Vision and Image Understanding*, 252:104301.

Xichen Zhang. 2023. Multimodal fake news analysis based on image–text similarity. *arXiv*.

A Source Channels Used for Dataset Construction

Table 5: Authentic and Fake News Sources (English)

Real Source	Fake Source
https://www.youtube.com/aljazeeraenglish	https://www.youtube.com/results?search_query=americannews.com
https://bangladeshpost.net/	https://www.youtube.com/@BusinessBasicsYT
https://islamchannel.tv/	https://www.youtube.com/@CipherSignal1
https://www.youtube.com/MiddleEastEye	https://www.youtube.com/@CrimeeWatchersss
https://www.nbcnews.com/	https://www.youtube.com/@DannyHaiphongYT
https://www.youtube.com/OwenJonesTalks	https://www.youtube.com/@Safety-Zone1
https://www.instagram.com/theeconomist/	https://www.youtube.com/@leftlens-us
https://www.facebook.com/thefrontlinebd/	—
https://www.facebook.com/thefrontpagebd/	—
https://www.trtworld.com/	—
https://www.youtube.com/bbcnews	—

Table 6: Authentic and Fake News Sources (Bangla)

Real Source	Fake Source
https://www.youtube.com/@BangladeshPratidinNews	https://www.youtube.com/@bangladeshnewsalert
https://www.youtube.com/@BusinessInspectionBD	https://www.youtube.com/@public360-m2o
https://www.youtube.com/@TheBusinessStandard	https://www.youtube.com/@PublicInsight-y7e
https://www.youtube.com/@Deshtvnews	https://www.youtube.com/@somoynews360
https://www.youtube.com/@EkusheyETV	https://www.youtube.com/@DannyHaiphongYT
https://www.youtube.com/@JamunaTVbd	https://www.youtube.com/@Somoy0334
https://www.youtube.com/@MaasrangaNewsbd	https://www.youtube.com/@leftlens-us