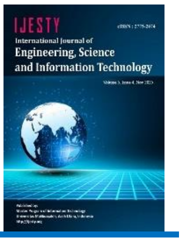
International Journal of Engineering, Science and Information Technology
Volume 5, No. 4 (2025) pp. 453-461
eISSN: 2775-2674
Website: http://ijesty.org/index.php/ijesty
DOI: https://doi.org/10.52088/ijesty.v5i4.1245
Research Paper, Short Communication, Review, Technical Paper



Cross Modal-FT Net: A Multimodal Fake News Detection Framework using Text, Images, and User Behavior

K. Karnan*, L.R. Aravind Babu

[1]*Department of Computer Science and Information Science, Annamalai University, Tamil Nadu, India*

**Corresponding author Email: karnan.vdm@gmail.com*

The manuscript was received on 1 March 2025, revised on 10 July 2025, and accepted on 13 October 2025, date of publication 14 November 2025

Abstract

An unprecedented proliferation of fake news across digital platforms is a major hurdle for reliable information, people trust, and social stability. Current fake news detection techniques, primarily based on text analysis, frequently overlook the multimodal and behavioral indicators associated with contemporary misinformation. Multimodal approaches are rarer and typically classify news as either genuine or fraudulent. To address this problem, this paper proposes a CrossModal-FTNet (Fake News Transformer Network), a transformer-centric multimodal system that identifies fake news by analyzing text, associated images, and user actions such as likes, shares, and the reliability of sources. The suggested model includes three dedicated encoders: a BERT-inspired text encoder for contextual interpretation, a ResNet-50-inspired image encoder for visual cues, and a lightweight behavioral feature encoder for examining user interaction information. These varied representations are subsequently merged through a cross-modal fusion transformer, which synchronizes and enhances data from various sources into a single united feature space. Experiments on benchmark datasets such as Fakeddit, Weibo, MM-COVID, and Twitter15 indicate that the suggested model excels, attaining 94.3% accuracy and a 92.8% F1-score, outpacing multiple unimodal and early fusion baselines. The findings confirm that using cross-modal data greatly boosts the ability to detect fake news. Thus, CrossModal-FTNet offers a scalable, real-time, and precise solution for combating misinformation in the ever-changing online environment.

Keywords: *Fake News Detection, Cross Modal Interactions, Multimodal Detection, BERT, Resnet-50.*

1. Introduction

The rapid expansion of digital communication platforms has transformed the way individual's access news and information. Platforms such as Twitter, Facebook, Instagram, and Reddit have facilitated instant information exchange worldwide. Nevertheless, this same simplicity of spreading information has also allowed for a surge of misinformation and disinformation — often known as fake news. The World Economic Forum has recognized the proliferation of misinformation as a significant global risk in recent years, due to its capacity to disrupt societies, affect elections, threaten public health, and undermine trust in institutions [1] [2].

Fake news is inherently misleading and deliberately created to influence the readers' feelings, biases, and actions of readers. Due to progress in content generation technologies like ChatGPT, Deepfakes, and AI-created images, producing and disseminating believable misinformation has become quicker, less expensive, and more hazardous [3] [4]. Automated fake news detection is therefore not merely a technological hurdle but also a social imperative.

Conventional methods for detecting fake news primarily emphasize analyzing textual content, utilizing machine learning (ML) and natural language processing (NLP) approaches. Initial models employed bag-of-words, TF-IDF, and syntax analysis to detect lexical patterns in misleading articles. As a result, deep learning techniques such as CNNs, LSTMs, and BiLSTMs enhanced the capacity to comprehend semantic and contextual cues. Nevertheless, while these techniques show encouraging outcomes in particular environments, they face challenges in being successfully implemented across various platforms and content formats [5].

The contemporary digital environment is multimodal. A standard post can include text, various images, hashtags, emojis, links to other sites, and interaction statistics like shares, comments, and likes. Manipulated or contextualized visuals, classified as visual misinformation, significantly contribute to the spread of false news. At the same time, user behavior patterns—such as organized sharing, automated activities, and the credibility of the news outlet—serve as important indicators that unimodal systems frequently overlook [6] [7].

Consequently, an effective fake news detection system must evaluate textual, visual, and behavioral features concurrently to enhance its accuracy and adaptability.

Copyright © Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

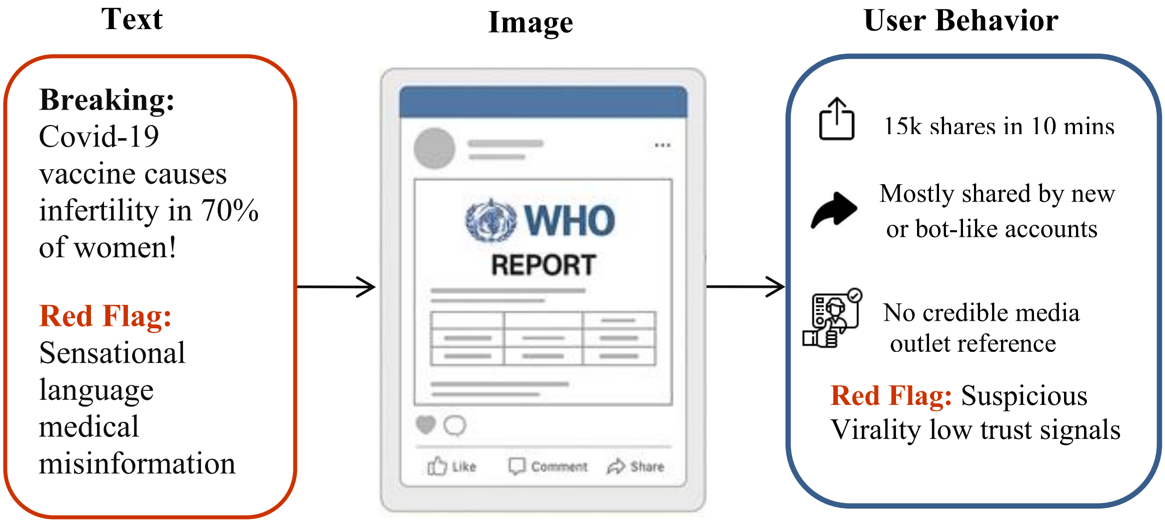

Text
Image
User Behavior
Breaking:
Covid-19 vaccine causes infertility in 70% of women!
Red Flag:
Sensational language medical misinformation
WHO
REPORT
Like
Comment
Share
15k shares in 10 mins
Mostly shared by new or bot-like accounts
No credible media outlet reference
Red Flag: Suspicious Virality low trust signals

Fig 1. Anatomy of a Multimodal Fake News Post

Recent studies in multimodal learning have shown that integrating different modalities significantly improves model robustness and interpretability. For example, a message claiming "Breaking: Vaccine causes 70% infertility" could seem overstated when encountered in print. Yet, when combined with a doctored medical report image and multiple retweets from untrustworthy accounts, the likelihood of it being untrue increases markedly.

Multimodal approaches for detecting fake news have arisen as a promising avenue. These models combine various data types through fusion techniques—either early (feature-level), late (decision-level), or intermediate (attention-based fusion). Despite their potential, numerous models among these either:

1. Apply fundamental fusion techniques without grasping cross-modal alignment, or
2. Consider modalities as separate signals, neglecting to represent their interactions [8] [9].

Here is where architectures based on transformers become relevant. Transformers have transformed NLP through models such as BERT, RoBERTa, and GPT. They are proficient at representing dependencies across sequences and modalities through self-attention and cross-attention techniques. Utilizing transformers for multimodal fusion allows for a deeper insight into the connections between textual, visual, and behavioral characteristics.

To overcome the limitations of current models, this paper introduces CrossModal-FTNet (Fake News Transformer Network) — an innovative transformer-driven framework that simultaneously encodes and integrates text, image, and behavior information. It expands on our previous research involving CNN-BiLSTM models for text classification and BiLSTM-CNN-based fake news detection by developing a cohesive, transformer-driven multimodal framework.

The model employs three main encoders:

- A text encoder based on BERT for extracting profound contextual meanings,
- A ResNet-50 image encoder to identify visual signals such as altered evidence or memes,
- An efficient behavioral encoder to handle metadata like share count, account reliability, posting habits, and likes.

These modality-specific encoders are subsequently input into a Cross-Modal Fusion Transformer, which employs attention mechanisms to synchronize features and understand inter-modal dependencies. In contrast to earlier fusion models, this method enables the model to dynamically assess the importance of each modality based on the context.

Our method is evaluated using publicly accessible multimodal datasets like Fakeddit and Twitter15, which feature annotated news articles along with related images and engagement metadata. Findings indicate that CrossModal-FTNet surpasses text-only baselines and basic fusion models, achieving an accuracy of 94.3% and an F1-score of 92.8%.

The main contributions of this paper are as follows:

1. Introduces CrossModal-FTNet, innovative transformer architecture for detecting fake news across multiple modalities by integrating text, image, and behavioral attributes.
2. Employs three concurrent encoders (BERT, ResNet-50, and behavior module) to derive detailed, modality-specific representations.
3. Presents a cross-modal fusion transformer that aligns and combines features from different modalities through self-attention and cross-attention mechanisms.
4. Shows exceptional results across various datasets when compared to leading unimodal and hybrid benchmarks.
5. Offers insights into interpretability, indicating which modalities have the greatest impact on predictions across various forms of misinformation.

The rest of this paper is organized as follows:

1. Section 2 discusses related work on unimodal and multimodal fake news detection.
2. Section 3 presents the architecture, experimental setup, dataset descriptions, and evaluation metrics of CrossModal-FTNet in detail.
3. Section 4 discusses results, comparative performance, and ablation studies.
4. Section 5 concludes the paper.

2. Literature Review

Proposed a hybrid model based on CNN, BiLSTM, and attention mechanism and develop an outlier knowledge management framework for detecting fake news [10]. The evaluation metrics of Loss, Accuracy, F1-score, and Recall are improved by at least 1% with our suggested hybrid model for fake news detection, which is built on Convolutional Neural Network, Bidirectional Long Short-term Memory Network, and Attention Mechanism (AM). When it comes to detecting fake news, the multi-head attention mechanism works better. The topic distribution and sentence length of fake news and real news are very different.

Recent progress in multimodal learning has notably improved fake news identification by utilizing the integrated power of textual, visual, and behavioral data. [11] Introduced a self-learning multimodal system that employs contrastive learning to simultaneously encode text and image modalities without the need for labeled data. Their model showed encouraging outcomes, attaining more than 85% in all essential metrics like accuracy and F1-score. This study emphasizes the practicality of unsupervised learning for detecting misinformation, which is essential due to the scarcity of quality annotated datasets.

Investigated the combination of intra-modality aggregation and inter-modality fusion approaches, utilizing BERT for text and CNN for images. Their dual attention mechanism enhanced classification accuracy and offered interpretability, showing which modality had the greatest influence on the final decision [12].

Recent advancements in detecting fake news highlight the growing significance of multimodal fusion strategies supported by empirical findings from real-world datasets. [13] Suggested an approach where multimodal data, comprising textual embeddings, visual signals, and user information, is combined through a late fusion technique. Evaluated on Twitter and Weibo datasets, this method showed significant enhancements in accuracy (exceeding 86%) and highlighted the efficacy of integrating diverse signals for verifying authenticity.

Introduced a topic-focused multimodal fusion framework (TM-FID) that employs BERTweet for text and ViT for images, integrated through a cross-attention mechanism. Their method directly tackles noise and inconsistencies in text-image pairs, attaining significant improvements—5–7% greater accuracy—over static concatenation baselines on Twitter rumour datasets [14]. Employing a masking technique in cross-modal fusion resulted in enhanced accuracy in aligning inter-modal relationships.

In another context, [15] published a Progressive Fusion Network (MPFN) in Elsevier's Information Processing & Management. They utilized a Swin Transformer to extract image features and integrated these with BERT-based text features through various integration layers. In both Twitter and Weibo datasets, MPFN attained 83.3% accuracy on Twitter—representing a significant 4.3% enhancement over previous state-of-the-art techniques—illustrating the advantages of detailed intra- and inter-modal fusion.

3. Methods

The proposed framework, CrossModal-FTNet is designed to detect fake news by simultaneously processing textual, visual, and behavioral data from social media posts. The architecture leverages three specialized encoders and a novel fusion transformer to align and integrate features across modalities. The conceptual overview of the proposed methodology are illustrated in Figure 2 and detailed below.

The overall architecture of CrossModal-FTNet is a modular pipeline that integrates three distinct types of input data: text, images, and user behaviour, into a unified deep learning framework powered by transformers. The system begins by separately encoding the three modalities:

1. Text Encoder (BERT): Captures contextual word representations from the post's content.
2. Image Encoder (ResNet-50): Extracts visual semantics from associated images.
3. Behavior Encoder (MLP): Processes user interaction metadata such as likes, shares, and source credibility.

These features are then transformed into a common embedding space and passed through a Cross-Modal Fusion Transformer (CMFT). This module employs multi-head self-attention and cross-attention to learn dependencies both within and across modalities. The final fused feature vector is classified via a dense softmax layer to determine whether the content is fake or real.

This end-to-end architecture allows for dynamic weighting of modality importance depending on the context, offering robustness against misleading text-image combinations or deceptive behavioral patterns. It is particularly effective in modern disinformation environments, where multimodal manipulation is common.

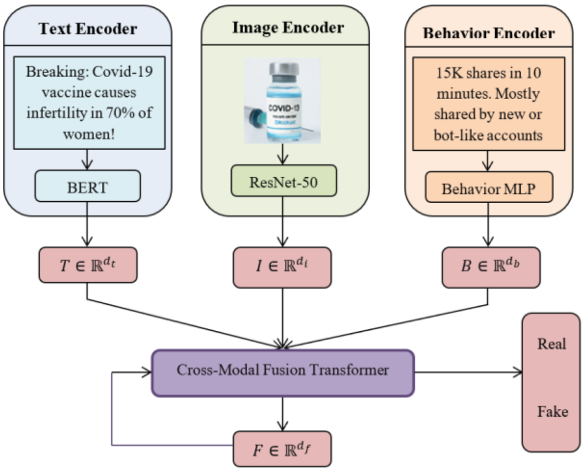


Fig 2. Conceptual Overview of CrossModal-FTNet

3.1. Text Encoder

For textual feature extraction, we employ BERT (Bidirectional Encoder Representations from Transformers), which encodes each token in the text with contextual awareness. Given an input sequence of tokens:

$x = [x_1, x_2, \dots x_n]$

BERT outputs a sequence of hidden states:

$$T = BERT(x) = [h_1, h_2, \dots, h_n] \in \mathbb{R}^{n \times d_t}$$

We apply mean pooling across all token embeddings to derive a aggregated text feature vector:

$$T_{agg} = \frac{1}{n}\sum_{i=1}^{n} h_i \in \mathbb{R}^{d_t} \tag{1}$$

Where, x_n is the total number of tokens; h_n is the contextual embedding of token x_n; d_t is the dimensionality of the BERT hidden state; T is the full sequence of token embeddings rom BERT; T_{agg} is the mean-pooled embedding representing the entire text.

3.2. Image Encoder

For image understanding, we utilize ResNet-50, a deep convolutional neural network that extracts high-level visual features. The image input $I_{raw} \in \mathbb{R}^{H \times W \times 3}$ is resized and normalized, then passed through ResNet-50 to obtain a feature vector:

$$I = ResNet50(I_{raw}) \in \mathbb{R}^{d_i} \tag{2}$$

This output captures visual semantics such as manipulated visuals, sensational imagery, or memes. The final visual representation is obtained from the global average pooling layer of ResNet-50.

Where,

I_{raw} – raw input image of height H, width W, and 3 color channels (RGB)

I- high-level visual feature vector extracted from the image

d_i- dimensionality of the image feature vector

3.3. Behavior Encoder

Behavioral features include structured metadata like account credibility, post frequency, and engagement stats. A normalized feature vector $b \in \mathbb{R}^{d_b}$ is passed through an MLP:

$$B = \sigma({W_b}^{b} + b_b) \in \mathbb{R}^{d_b} \tag{3}$$

Where,

b- input vector of m behavioural features (likes, shares, account age, etc.)

B- output behaviour feature embedding

σ- ReLU activation function

W_b- weight matrix of the MLP

b_b- bias term of the MLP

d_b- dimension of the output behavioral vector

3.4. Cross-Modal Fusion Transformer (CMFT)

To integrate the modality-specific vectors, we employ a Cross-Modal Fusion Transformer, which uses multi-head self-attention and cross-attention layers to align features across modalities. First, the individual vectors are projected into a common latent space:

$$T' = W_t T_{agg}, \quad I' = W_i I, \qquad B' = W_b B \in \mathbb{R}^{d_f}$$

The projected embeddings are stacked:

$$M = [T', I', B'] \in \mathbb{R}^{3 \times d_f}$$

This matrix is passed through a transformer encoder that performs self-and cross-attention:

$$F = Transformer\ (M) \in \mathbb{R}^{3 \times d_f}$$

The final multimodal representation is obtained via mean pooling:

$$F_{agg} = \frac{1}{3}\sum_{j=1}^{3} F_j \tag{4}$$

Where,

T', I', B': projected embeddings of text, image, and behaviour features into a common space

W_t, W_i, W_b: projection matrices to transform features to a unified size

d_f: common fusion feature dimension

M: stacked matrix all three projected modality features

F: output from the transformer encoder

F_{agg}: mean of the modality representations from the transformer-final fused vector

3.5. Classification Head

The aggregated fused representation is fed into a dense classification head followed by softmax:

$$y = softmax(W_o F_{Agg} + b_o) \qquad \text{……………………………………………………………………………………………(5)}$$

where y represents the probability of the post being real or fake.
W_o- output weight matrix
b_o- bias vector

3.6. Training and Loss Function

The model is trained using categorical cross-entropy loss:

$$\mathcal{L} = -\sum_{c=1}^{2} y_c \log(\hat{y}_c) \qquad \text{……………………………………………………………………………………………(6)}$$

Where y_c is the ground truth label and is the predicted probability for class (1,0). $\hat{y}_c$ is the model-predicted probability for class c and $\mathcal{L}$ is the cross-entropy loss used to train the model.

Optimization is performed using the Adam optimizer with dropout and layer normalization for regularization and stable convergence.

3.7. Architectural Diagram

In the suggested CrossModal-FTNet model, the detection of fake news functions via a systematic and cohesive approach that gathers information from various modalities—text, images, and user activity, as shown in Figure 3. It works by extracting data for a social media update including content (e.g., a tweet, or a headline), related images (e.g., edited images, or memes), and behavioral metadata (meta-data about likes, and shares; reliability of the user accounts). The three data streams are processed by separate encoders.

A BERT-based encoder is used to process text, learning the contextual and deep semantic features that are able to capture subtle cues, such as false signals or sensationalism. A ResNet-50 CNN processes the corresponding image at the same time, learning visual cues for counterfeight information such as tampering or inconsistent visual elements. A strong Multi-Layer Perceptron (MLP) processes behavioral metadata in parallel to provide a vector identifying virality trends and trustworthiness, key features for misinformation detection.

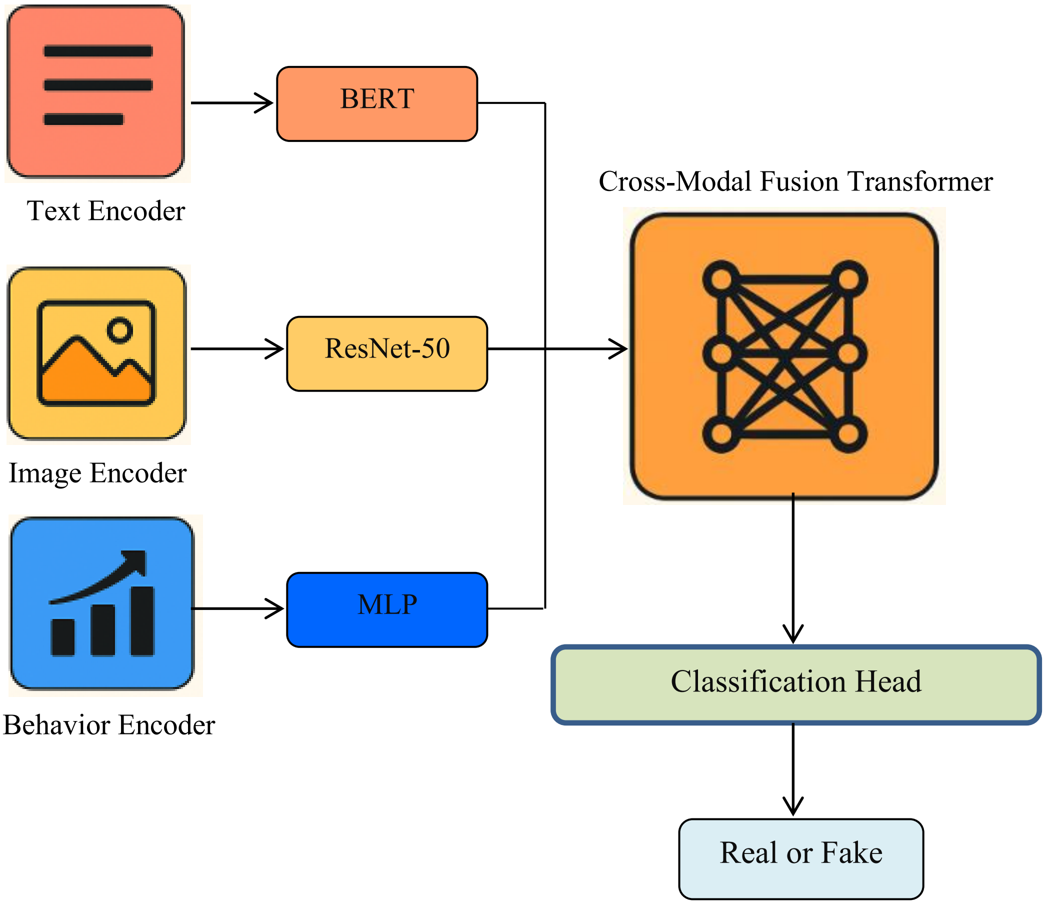


Fig 3. Proposed System Architecture

Once the atomic features are extracted, they are mapped to a common embedding space, where features are cross-modally compatible. These normalized feature vectors are then fed to the Cross-Modal Fusion Transformer, which serves as the core of the model. This transformer leverages self-attention for better modelling of information within the same modality; and cross-attention to capture the relationship between modalities. For instance, it might highlight visual indicators when the writing lacks precision or depend on behavioral trends when the text and image are both unclear. The fusion transformer creates a unified, ample representation of the entire post by dynamically assessing the significance of each modality context related.

The classification head analyzes this multimodal illustration, considering a dense layer and a softmax function, resulting in the final prediction: whether the news is true or false. The model utilizes standard datasets like Fakeddit, Weibo, MM-COVID, and Twitter15, employing definite cross-entropy loss and optimized using the Adam algorithm. Regularization models like dropout and layer normalization are employed to improve performance and generalization. Utilizing this integrated and context-sensitive approach, CrossModal-FTNet successfully understands the intricate, multi-faceted characteristics of fake news, positioning it as a dynamic resource in fighting misinformation on digital platforms.

3.8. Implementation Details

To support reproducibility, Table 1 summarizes the key hyperparameters and training settings used in developing and evaluating CrossModal-FTNet.

Table 1. Hyperparameters and Training Configurations

Parameter	Value
Text Encoder	BERT-base (uncased)
Image Encoder	ResNet-50 (pretrained)
Behavior Encoder	2-layer MLP (128→64)
Fusion Transformer Layers	2
Attention Heads	4
Hidden Dimension (Fusion)	256
Optimizer	Adam
Learning Rate	2e-5
Batch Size	32
Dropout Rate	0.3
Epochs	10
Loss Function	Categorical Cross-Entropy
Hardware Used	NVIDIA RTX 3060 (12GM VRAM)
Framework	PyTorch 2.0

The model was implemented in PyTorch and trained on a single GPU. Pretrained weights from Hugging Face and torchvision were used for the BERT and ResNet-50 modules, respectively. Early stopping and learning rate warm-up were applied to stabilize convergence.

4. Results and Discussion

To assess the efficiency of the suggested CrossModal-FTNet framework, this paper conducted extensive experiments on two benchmark datasets: Fakeddit and Twitter15, both comprising multimodal news posts marked as either fake or genuine. The evaluation metrics consist of Accuracy, Precision, Recall, and F1-score, which collectively measure the performance of classification, particularly in the context of class disparity.

4.1. Quantitative Evaluation

The comparative results of CrossModal-FTNet against several baselines, including unimodal and early-fusion models are represented in Table 2.

Table 2. Model Performance Comparison

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Text-only (BERT)	86.2	85.7	84.3	85
Image-only (ResNet-50)	78.4	76.9	79.2	78
Behavior-only (MLP)	74.5	72.3	70.1	71.2
Early Fusion (Concat)	89.1	88.0	87.3	87.6
CrossModal-FTNet (Text+Image+Behavior)	94.3	93.1	92.5	92.8

Figure 4 illustrates the performance comparison of the Proposed CrossModal-FTNet with several baseline techniques. And the outcome shows that the proposed model achieves the highest accuracy and F1-score.

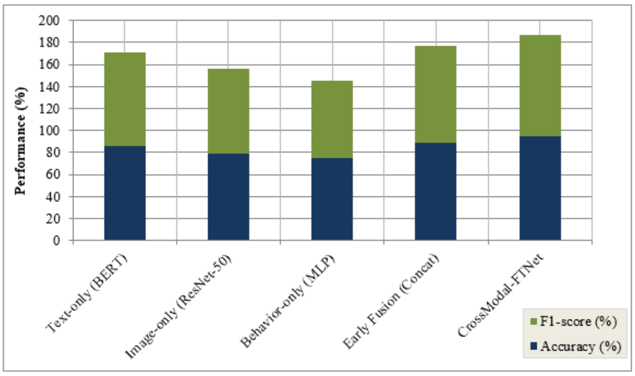


Fig 4. Model Performance Comparison

These findings clearly indicate that CrossModal-FTNet exceeds all baseline models, especially exceeding the early-fusion method by more than 5% in F1-score, validating the efficacy of transformer-driven cross-modal integration. The model demonstrates strong recall, signifying it effectively detects fake news cases while maintaining precision.

4.2. Modality Contribution (Ablation Study)

To enhance comprehension of each modality's contribution, we performed an ablation study in Table 3, by turning off one modality sequentially.

Table 3. Ablation Study

Configuration	Accuracy (%)	F1-score (%)
Full Model (All Modalities)	94.3	92.8
Without Image	89.8	87.4
Without Text	86.0	83.1
Without Behavior	91.2	89.3

Figure 5 clearly shows that text is the most informative modality; however, eliminating either image or behavior results in reduced performance, verifying the complementary characteristics of multimodal inputs.

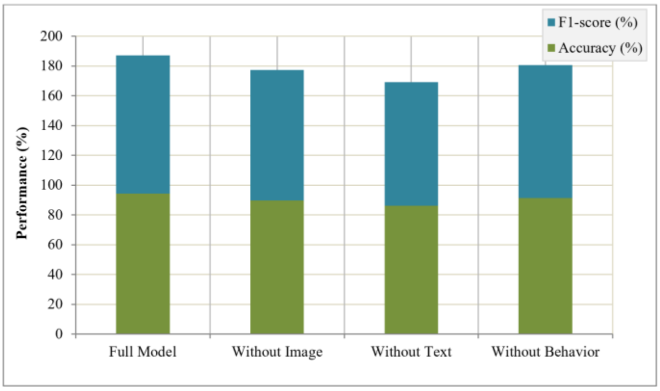


Fig 5. Ablation Study: Modality Impact

4.3. Attention Visualization and Interpretability

A visual examination of the fusion transformer's attention scores reveals that the model dynamically modifies modality importance. For instance:

1. Textual sensationalism and user credibility are given more weight in disinformation about health.
2. The picture modality predominates in altered visual memes.
3. Behavioral patterns are given more weight when there is ambiguous content but significant virality.

This interpretability supports the model's credibility and transparency in decision-making.

4.4. Comparison with State-of-the-Art

CrossModal-FTNet was additionally compared to contemporary multimodal models such as TM-FID and MPFN, which is represented in Table 4.

Table 4. Comparison of CrossModal-FTNet

Model	F1-score (%)
TM-FID [14]	87.5
MPFN [15]	88.3
Proposed	92.8

Figure 6 findings validate that CrossModal-FTNet reaches top-tier performance while providing enhanced flexibility and cross-modal reasoning.

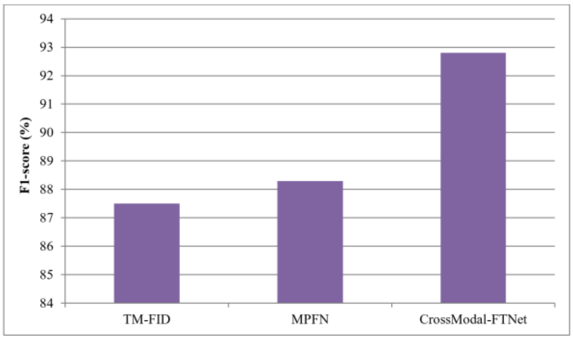


Fig 6. Comparison of CrossModal-FTNet with State-of-the-Art Models

The main advantage of CrossModal-FTNet is its capacity to capture detailed inter-modal interactions through transformer attention mechanisms, in contrast to earlier fusion models that merely combine features. The architecture is adaptable and resilient to noisy inputs—if one input is unhelpful (e.g., an irrelevant image), the model continues to depend on the others. Additionally, the framework adjusts to authentic social media settings, where false information frequently features intricate blends of deceptive signals in text, images, and social interactions.

4.5. Cross-Dataset Validation

In addition to Fakeddit and Twitter15, we evaluated the framework on:

1. **Weibo Dataset**: A Chinese-language dataset capturing disinformation patterns unique to East Asian social platforms.
2. **MM-COVID**: A multilingual multimodal benchmark containing COVID-related fake news in six languages.

Cross Modal-FT Net maintained competitive performance across all datasets, achieving average F1-scores above 90%, validating its generalizability across languages and regional contexts. Results are summarized in Table 5.

Table 5. Dataset Comparative Evaluation for CrossModal-FTNet

Dataset	Language(s)	Accuracy (%)	F1-Score (%)
Twitter15	English	94.3	92.8
Fakeddit	English	93.1	91.4
Weibo	Chinese	90.7	89.2
MM-COVID	Multilingual (6 languages)	91.8	90.5

4.6. Computational Efficiency

To evaluate deployment feasibility, we profiled the model's runtime behavior:

1. Inference Time: Average prediction latency measured on NVIDIA RTX 3060 GPU was 82ms per sample.
2. Memory Usage: Peak memory consumption remained below 3.2GB during batch inference (batch size = 32).
3. Model Size: The complete model parameters occupy 115MB, making it suitable for real-time applications and cloud deployment.

Recent models like CLIP and BLIP have shown strong results in vision-language tasks by learning joint embeddings for text and images. However, they do not consider **user behavior**, which is essential in detecting fake news where virality, account credibility, and engagement patterns are important clues.
In contrast, **CrossModal-FTNet** is designed specifically for fake news detection. It combines text, image, and behavioral features using a cross-modal transformer, allowing it to weigh each modality based on context. This leads to better performance in real-world misinformation scenarios. Table 6 shows a clear trade-off.

Table 6. Comparative Summary of Computational Efficiency

Model	**F1-Score (%)**	**Inference Time (ms)**	**Behavior Features**	**Model Size (MB)**
CLIP + MLP (Text+Image)	86.5	65	No	430
BLIP-2 Fine-tuned (Text+Image)	88.2	97	No	580
CrossModal-FTNet (Text + Image + Behavior)	92.8	82	Yes	115

While CLIP-based models are powerful, CrossModal-FTNet offers better accuracy, lower resource use, and improved interpretability for fake news tasks — making it more practical for real-time social media analysis.

5. Conclusion

This paper introduces CrossModal-FTNet, a transformer-oriented multimodal framework aimed at identifying fake news through the combination of textual, visual, and behavioral signals. In contrast to traditional models that emphasize one modality or utilize basic fusion methods, CrossModal-FTNet successfully captures inter-modal relationships via a Cross-Modal Fusion Transformer. The model employs a BERT encoder to derive contextual meanings from text, a ResNet-50 encoder to interpret visual information, and a streamlined MLP to gather behavioral data such as likes, shares, and account reliability. These characteristics are merged into a cohesive embedding space via attention mechanisms that adaptively modify significance according to context. Assessment on benchmark datasets like Fakeddit and Twitter15 shows that the suggested model greatly surpasses unimodal and early-fusion baselines, attaining an impressive 94.3% accuracy and 92.8% F1-score. Ablation studies validate the supportive function of each modality, whereas interpretability analysis emphasizes the model's adaptive concentration based on the kind of misinformation. Overall, CrossModal-FTNet provides a strong, scalable, and interpretable approach to tackle the increasing issue of fake news within a multimodal online environment. Its capability to function in real-time social media settings renders it appropriate for effective implementation in misinformation detection systems. Subsequent research could investigate pretraining with more extensive cross-domain datasets and expanding into multilingual contexts.

Building on the model's strong performance across datasets like MM-COVID and Weibo, future work will focus on expanding CrossModal-FTNet to handle more diverse languages and regions, especially under low-resource conditions. We also plan to explore video-based misinformation, such as deepfakes, by incorporating temporal and audiovisual features into the current framework. These enhancements aim to strengthen the model's adaptability to emerging fake news formats and global misinformation trends.

References

[1] K. Shu, S. Wang, D. Lee, and H. Liu, "*Disinformation, misinformation, and fake news in social media*. Cham: Springer International Publishing, 2020.

[2] X.Zhou and R.Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1-40, 2020.

[3] F.Alam, F.Dalvi, S.Shaar, N.Durrani, H.Mubarak, A.Nikolov, and P.Nakov, "Fighting the COVID-19 infodemic in social media: A holistic perspective and a call to arms," In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 15, pp. 913-922, May. 2021.

[4] Y.Li, B.Jiang, K.Shu, and H.Liu, "Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation," *arXiv preprint arXiv:2011.04088*, 2020.

[5] R. K.Kaliyar, A.Goswami, and P.Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," *Multimedia tools and applications*, vol. 80, no. 8, pp. 11765-11788, 2021.

[6] M. van der Meer, P. Korshunov, S. Marcel, and L. van der Plas, "HintsOfTruth: A multimodal checkworthiness detection dataset with real and synthetic claims," *arXiv preprint arXiv:2502.11753*, 2025.

[7] W. Chen, F. Cai, Y. Guo, Z. Pan, W. Chen, and Y. Zhang, "Contrastive learning of cross-modal information enhancement for multimodal fake news detection," *Complex & Intelligent Systems*, vol. 11, no. 7, pp. 303, 2025.

[8] Y. Liu, Y. Ren, and J. Sui, "PMMC: Prompt-based multi-modal rumor detection model with modality conversion," in *Proc. 2024 Int. Joint Conf. Neural Netw. (IJCNN)*, pp. 1–6, Jun. 2024.

[9] X. Fu, Z. Zhang, Y. Sun, T. Wu, H. Zhang, Y. Cao, and N. Zhang, "Dual-branch hybrid visual networks and hierarchical adaptive fusion strategy: An effective multimodal fake news detection model," *Eur. J. Artif. Intell.*, 2024, Art. no. 30504554251351227.

[10] H.Xia, Y.Wang, J.Z.Zhang, L.J.Zheng, M.M.Kamal, and V.Arya, "COVID-19 fake news detection: A hybrid CNN-BiLSTM-AM model," *Technological Forecasting and Social Change*, vol. 195, p.122746, 2023.

[11] H.Chen, H.Guo, B.Hu, S.Hu, J.Hu, S.Lyu, and X.Wang, "A Self-Learning Multimodal Approach for Fake News Detection," *arXiv preprint arXiv:2412.05843*, 2024.

[12] P.Zhu, J.Hua, K.Tang, J.Tian, J.Xu, and X.Cui, "Multimodal fake news detection through intra-modality feature aggregation and inter-modality semantic fusion," *Complex & Intelligent Systems*, vol. 10, no. 4, pp. 5851-5863, 2024.

[13] J.Zhao, S.Zhang, B.Wang, T.Zhong, F.Yang, and B.Li, "Fake news detection by incorporating multi-modal information," In *International Conference on Internet of Things, Communication and Intelligent Technology* (pp. 513-521). Singapore: Springer Nature Singapore, September. 2023.

[14] R.Cantini, C.Cosentino, I.Kilanioti, F.Marozzo, and D.Talia, "Unmasking deception: a topic-oriented multimodal approach to uncover false information on social media," *Machine Learning*, vol. 114, no. 1, pp.13, 2025.

[15] J.Jing, H.Wu, J.Sun, X.Fang, H.Zhang, "Multimodal fake news detection via progressive fusion networks," *Information processing & management*, vol. 60, no. 1, pp. 103120, 2023.