Innovations in Computing and Emerging Technologies

Print ISSN: 3106-6518 | Online ISSN: 3106-6526
Issue: 2026 (01) | *Peer-Reviewed Academic Journal*

Research Article

Cross-Modal Inconsistency Detection with Pooled Fusion: A Controlled Study for Multimodal Fake News

Waqas Gulzar Gondal[1]
Sajid Ullah Khan[1]
Atlas Gondal[2]

Article Information:

Article No.	ICET-20260101-0421
Submission Date	January 01, 2026
Acceptance Date	April 30, 2026
Publication Date	June 19, 2026

Affiliations:

[1] Department of Computer Science, University of Lahore, Lahore, Pakistan
[2] College of Computer and Information Science, Princess Nourah Bint Abdulrahman University, Riyadh, Kingdom of Saudi Arabia (KSA).

Innovations in Computing and Emerging Technologies, 2026 (01)
DOI: TBD

Abstract

Fake news often pairs misleading text with authentic images, so cross-modal consistency can be a useful signal for multimodal detection. In this paper, we test whether a lightweight pooled attention fusion layer—operating on pooled robustly optimised bidirectional encoder representations from transformers model and vision transformer model embeddings—provides consistent improvements over strong fusion baselines on binary Fakeddit classification. Using stratified subsets from the official splits (training set/validation set/testing set = 50,000/10,000/10,000) and three random seeds (mean ± standard deviation), multimodal fusion improves over unimodal models, but pooled attention does not consistently outperform late fusion or gated fusion. On the primary setting, late fusion achieves the best mean accuracy (88.78%), gated fusion the best mean area under the receiver operating characteristic curve (0.9516), and pooled attention remains competitive (88.50% accuracy; 0.9460 area under the receiver operating characteristic curve). Overall, pooled attention is a useful coarse interaction term, but it is not sufficient evidence of fine-grained inconsistency reasoning.

Keywords: Keywords: Multimodal Fake News Detection, Cross-Modal Consistency, Multimodal Fusion, Vision Transformer (ViT), RoBERTa, Misinformation Detection

1 Introduction

The rapid spread of misinformation on social platforms has made fake news detection (FND) an important problem for Natural Language Processing (NLP), computer vision, and multimodal learning [1] [2] [3], [4], [5], [6]. In multimodal settings, posts often pair short textual claims with images that may be authentic but misleading in context. In such cases, the key signal is not only the text or image alone, but whether the two modalities support each other.

In this paper, we address a focused and practically relevant question: under controlled, multi-seed evaluation, does adding a lightweight pooled-attention interaction between strong unimodal encoders yield measurable gains over simpler fusion baselines?

This question aligns with cross-modal inconsistency detection, which models whether a textual claim is semantically compatible with the accompanying image [7], [8], [9]. Prior multimodal FND systems typically encode text and images separately and combine them using concatenation, pooling, or attention-based fusion [10], [11], [12], [13]. However, attention is sometimes interpreted as evidence of fine-grained reasoning even when it operates only on pooled representations. Accordingly, we treat pooled attention as a coarse interaction module and align our claims with this granularity. To isolate the effect of fusion, we compare against strong unimodal and simple fusion baselines under matched data splits, optimisation settings, and a multi-seed protocol.

This distinction matters in our setting. We use Robustly Optimised BERT Pretraining Approach (RoBERTa) and Vision Transformer (ViT) encoders, but the fusion block applies attention only between global pooled embeddings (RoBERTa first-token/[CLS]-style text embedding and the ViT [CLS] image embedding). We therefore treat the method as an architecturally simple coarse interaction module rather than a patch-level or token-level inconsistency detector. Pooled attention fusion is nevertheless common in practical multimodal pipelines because it is simple, stable to train, and computationally efficient; we therefore evaluate whether even this limited interaction term provides measurable benefits beyond strong unimodal and simple fusion baselines.

We conduct our primary experiments on a larger Fakeddit subset with 50,000/10,000/10,000 train/validation/test samples, and use a 5,000/1,000/1,000 subset as a smaller preliminary setting. Across both settings, multimodal fusion improves over unimodal baselines, while pooled attention fusion is competitive but not consistently superior to late fusion or gated fusion. These results help narrow what can be claimed about global pooled attention mechanisms in multimodal FND.

The main contributions of this paper are threefold:

- Problem framing. We study multimodal fake news detection through the lens of claim–image consistency, while explicitly separating this motivation from the more limited capabilities of pooled-embedding fusion.
- Controlled model comparison. We evaluate a RoBERTa–ViT classifier with a single-block pooled attention fusion module against text-only, image-only, late-fusion, and gated-fusion baselines under the same data splits, optimisation setup, and multi-seed protocol.
- Empirical finding. We show that multimodal fusion is helpful on Fakeddit, but pooled attention fusion is competitive rather than dominant, indicating that stronger claims about inconsis-

Published on June 19, 2026

tency reasoning require finer-grained fusion and stronger empirical validation.

By aligning the paper's claims with what the current implementation actually does, we aim to provide a defensible and reproducible baseline for future work on fine-grained multimodal misinformation detection.

2 Related Work

Research on fake news detection spans text-based modelling, multimodal fusion, and context-aware reasoning. Recent surveys and overviews report two broad observations: transformer backbones now dominate the field, yet reliable multimodal reasoning remains difficult, particularly when text and images are only loosely aligned [3], [14], [15].

2.1 Content-Based Fake News Detection

Early FND systems relied on hand-crafted linguistic features and classical classifiers, whereas current approaches predominantly use deep encoders, especially transformer-based language models [9], [11], [16], [17], [18], [19]. These methods remain strong when the textual content carries most of the signal, but they do not address whether an accompanying image supports, weakens, or contradicts the claim. However, many real-world posts are multimodal, and incorporating visual evidence becomes important when the image changes the interpretation of the text.

2.2 Multimodal Fake News Detection

Multimodal FND models typically encode text and image streams separately and then combine them through concatenation, bilinear pooling, hierarchical fusion, or attention-based modules [10], [12], [20], [21], [22], [23]. More recent work increasingly adopts transformer-based encoders together with cross-modal interaction layers [13], [24], [25]. Even so, attention is sometimes treated as a generic fusion improvement, with limited distinction between coarse-pooled interaction and genuinely fine-grained token-to-patch reasoning.

In practice, many pipelines favour pooled encoders with simple fusion because they are easy to implement, computationally efficient, and stable to train – a pragmatic choice for title – image datasets such as Fakeddit, where the pairing can be weakly aligned and the supervision is based on post labels rather than explicit inconsistency annotations. This naturally leads to inconsistency-aware modelling, where the objective is not only to fuse modalities, but also to capture whether they agree semantically.

2.3 Cross-Modal Inconsistency as a Detection Signal

Several studies argue that semantic mismatch between modalities is an important cue for misinformation detection. Chen et al. model modality ambiguity, while MFIR introduces dedicated inconsistency-learning components for explainable multimodal detection [7], [26]. Related analyses further show that weak labels, contextual variation, and noisy multimodal pairings can make simple similarity-based assumptions unreliable [8], [9]. Taken together, this literature supports the motivation for claim–image consistency modelling, but it also makes clear that stronger inconsistency claims should be reserved for models that truly operate at a finer granularity.

However, controlled large-scale evidence remains limited on whether pooled attention fusion – a coarse interaction design – yields consistent gains over strong simple fusion baselines under multi-seed evaluation.

2.4 Contextual and Graph-Based Approaches

Another major direction augments content features with propagation, user, or graph-structured signals [27], [28], [29]. These methods often improve robustness and generalisation, but they address a different problem from the one studied here: whether content-only multimodal fusion can recover useful interaction cues from a title–image pair.

In summary, prior work shows that multimodal fusion is useful and that inconsistency-aware modelling is promising, especially for weakly aligned claim–image pairs. What remains less clear is whether many reported "attention-based" multimodal detectors perform genuinely fine-grained cross-modal reasoning or instead add a coarse interaction term on top of strong unimodal encoders. We address this narrower gap by evaluating a pooled-embedding RoBERTa–ViT attention fusion model at 50k scale under controlled splits and a multi-seed protocol, and by comparing it directly against strong simple baselines (late fusion and gated fusion).

3 Methodology

In this section, we describe the multimodal fake news detection task under a claim–image consistency motivation, the architecture of our proposed model, and the training procedure. Our goal is to provide a clear

Published on June 19, 2026

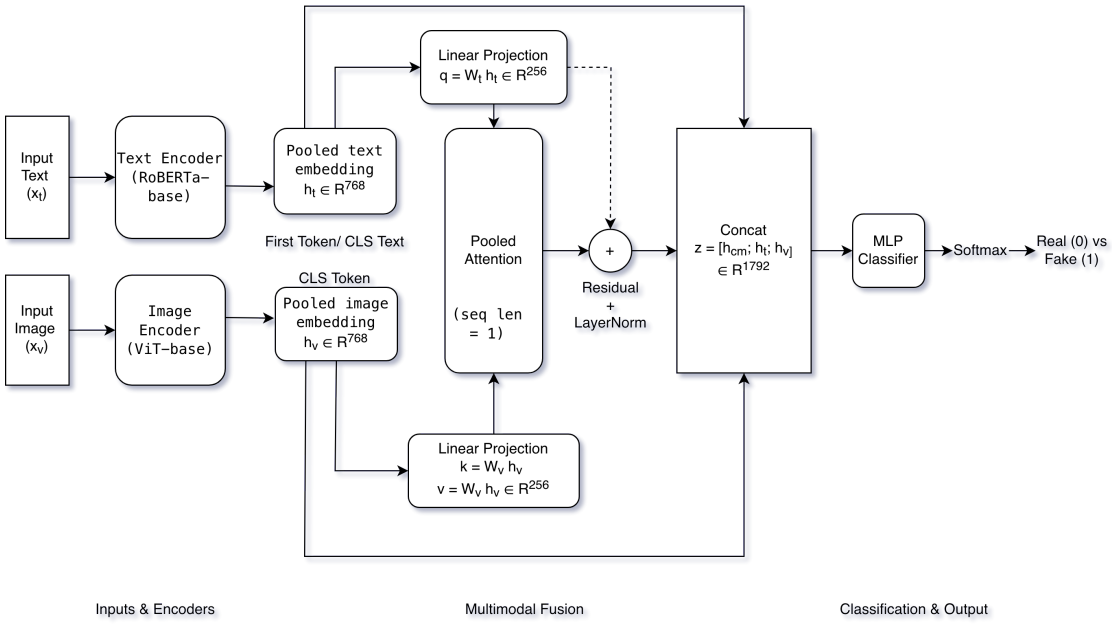

Figure 1: Overview of the proposed architecture with tensor dimensions

and reproducible description that closely matches the implemented system.

3.1 Problem Formulation

We consider the setting where each news item is represented as a pair (Xt, Xv), where Xt is the textual component (e.g., a news title or short caption) and Xv is the associated image. The task is to predict a binary label y{0,1}, where y=0 denotes a real post and y=1 denotes a fake post.

Our motivation is that fake posts may contain mismatches between the textual claim and the associated image. However, the dataset label is not a direct annotation of text–image inconsistency. We therefore treat cross-modal consistency as a latent cue that may help classification, rather than as a supervised target in its own right [7], [8], [9].

Formally, our model defines the conditional distribution in Eq. 1, where θ denotes all trainable parameters:

$$p_\theta(y \mid x_t, x_v) \tag{1}$$

The model is trained to minimise the standard cross-entropy loss over labelled text–image pairs. The key design choice lies in how we encode each modality and how we fuse them in a way that makes inconsistencies salient to the classifier.

3.2 Overview of the Proposed Architecture

Our architecture, illustrated at a high level in Fig. 1, consists of three main components:

- A text encoder, based on RoBERTa-base, that produces a compact representation of the input text;
- An image encoder, based on ViT-base-patch16-224, that produces a compact representation of the input image; and
- A pooled attention fusion module that projects the pooled text and image embeddings into a shared hidden space, applies a single-step attention interaction, and passes the fused representation together with the original unimodal embeddings to a classifier.

This design is inspired by prior work demonstrating the effectiveness of transformer-based encoders for both text and images [11], [19], [24], as well as multimodal FND models that rely on attention-based fusion [7], [10], [12]. The important limitation is that our fusion operates only on pooled embeddings. As a result, the model captures a coarse interaction between modalities, not token-level or patch-level contradiction patterns.

3.3 Text Encoder

For the textual modality, we use RoBERTa-base as our encoder, which offers strong performance on downstream NLP tasks and is a common baseline in modern

Published on June 19, 2026

NLP. We initialise it with pretrained weights (roberta-base). RoBERTa is a robustly optimised BERT-style pretraining approach with improvements in both training data and hyperparameters, and it often improves upon standard BERT on downstream tasks.

Given a tokenised input sequence of length L, RoBERTa produces a sequence of contextualised embeddings as shown in Eq. 2, where dt=768 is the hidden size:

$$H_t = [h_{t,1}, h_{t,2}, \ldots, h_{t,L}] \in R^{L \times d_t} \tag{2}$$

Following common practice in classification tasks, we take the embedding corresponding to the first token (analogous to the [CLS] token in BERT) as a summary representation of the text, as in Eq. 3:

$$h_t = h_{t,1} \in R^{d_t} \tag{3}$$

In our implementation, we use the pretrained weights from Hugging Face and fine-tune the text encoder end-to-end during training to allow the model to adapt to the specific FND task. This approach is consistent with recent observations in multimodal FND that fine-tuning transformer encoders yields substantial performance gains [11], [19].

3.4 Image Encoder

For the visual modality xv, we adopt ViT-Base/16 (google/vit – base – patch16 – 224) as our image encoder. ViT has shown strong performance across a wide range of CV tasks by applying the transformer architecture, originally developed for NLP, directly to sequences of image patches.

Each input image is first resized and normalised to a fixed resolution (224×224). The ViT model splits the image into non-overlapping patches, embeds each patch, and adds positional encodings, yielding a sequence of patch embeddings. A special classification token is prepended to this sequence. After passing through multiple transformer layers, ViT outputs the sequence in Eq. 4, where P is the number of patches and dv=768 is the hidden size:

$$H_v = [h_{v,\text{cls}}, h_{v,1}, \ldots, h_{v,P}] \in R^{(P+1) \times d_v} \tag{4}$$

We use the embedding of the classification token as the global image representation, shown in Eq. 5:

$$h_v = h_{v,\text{cls}} \in R^{d_v} \tag{5}$$

Similar to the text encoder, we use pretrained ViT weights and fine-tune the image encoder end-to-end during training. This setup is consistent with recent multimodal FND studies that leverage pretrained vision backbones with fine-tuning to achieve strong performance [10], [12], [21].

3.5 Pooled Attention Fusion

The core of our architecture is an architecturally simple fusion module applied to pooled text and image representations. Rather than using simple concatenation alone, we introduce a single-step attention interaction in which the text embedding acts as a query over the image embedding, which serves as both key and value. This follows the general template of cross-attention in vision–language models, but in a deliberately compact pooled form tailored to binary classification.

Concretely, we first project the text and image embeddings into a shared hidden space of dimension dh using the projections in Eq. 6, where Wt and Wv are learnable matrices in R{dh × 768} :

$$q = W_t h_t, \quad k = W_v h_v, \quad v = W_v h_v \tag{6}$$

For attention, we treat these projected vectors as single-token sequences, i.e., q,k v R{1xdh}. We then apply attention as in Eq. 7:

$$\text{Attn}(q,k,v) = \text{softmax}\left(\frac{qk^T}{\sqrt{d_k}}\right) v \tag{7}$$

Here dk denotes the key/query dimension in the scaled dot-product term. For readability, ([eq:attention]) shows the single-head scaled dot-product form; in implementation we use PyTorch's MultiheadAttention with a batch-first interface.

Because both the text and image sides are treated as sequences of length 1, qkT is a 1×1 scalar and the softmax is taken over a single element (hence equals 1). Consequently, this block behaves as a learned global interaction/mixing layer between pooled embeddings rather than a token-to-patch alignment mechanism. The attention output is then combined with the original text projection via a residual connection and layer normalisation, as in Eq. 8:

$$\tilde{h}_{cm} = \text{LayerNorm}\big(q + \text{Dropout}(\text{Attn}(q,k,v))\big) \tag{8}$$

This yields a fused cross-modal representation h{cm}R{dh }.

In implementation terms, each sample follows the tensor flow in Eq. 9:

$$\begin{aligned} & h_t \in R^{768},\ h_v \in R^{768} \\ & \to q,k,v \in R^{256} \\ & \to \tilde{h}_{cm} \in R^{256} \\ & \to z = [\tilde{h}_{cm}; h_t; h_v] \in R^{1792} \end{aligned} \tag{9}$$

Accordingly, we interpret this module as coarse cross-modal reweighting in a shared space. It can model whether the global text and image summaries interact in a way that is useful for classification, but it cannot localise which words conflict with which image regions. Despite its limited granularity, this block can

Published on June 19, 2026

still learn a structured shared-space transformation that complements direct concatenation before classification.

3.6 Gated Fusion Baseline (Learned Modality Weighting)

In addition to concatenation and cross-attention, we evaluate a gated fusion baseline that learns to weight the contribution of each modality. Given pooled unimodal embeddings ht and hv, we compute a scalar gate α(0,1) using Eq. 10, where Wg and bg are trainable gate parameters:

$$\alpha = \sigma\big(w_g^T[h_t; h_v] + b_g\big) \tag{10}$$

We then form the fused representation in Eq. 11:

$$h_{\text{gate}} = \alpha h_t + (1-\alpha)h_v \tag{11}$$

We then use the same classifier head (MLP) on top of the fused representation (and, when applicable, the unimodal embeddings) so that the comparison isolates the effect of learned modality weighting as a simple alternative fusion mechanism.

3.7 Classification Head

After computing the cross-modal representation hcm, we combine it with the original unimodal embeddings to preserve both the interaction term and the raw modality-specific information. This design is motivated by the strength of the unimodal signals in Fakeddit and by the limited granularity of the pooled fusion block. Using hcm alone would compress away information that remains predictive on this dataset. We therefore form the concatenated vector in Eq. 12:

$$z = [h_{cm}; h_t; h_v] \in R^{d_h+d_t+d_v} \tag{12}$$

We then pass through a small feed-forward network as shown in Eq. 13, where Wf and bf are trainable parameters:

$$z' = \text{LayerNorm}\big(\text{Dropout}(\text{ReLU}(W_f z + b_f))\big) \tag{13}$$

This is followed by a final linear layer that outputs logits for the two classes, as in Eq. 14, with trainable Wc and bc:

$$\ell = W_c z' + b_c \tag{14}$$

The predicted probabilities are obtained via a softmax over. We train this classifier end-to-end together with the pooled attention fusion and projection layers, as well as the text and image encoders.

3.8 Training Objective and Optimisation

We train the model using the standard cross-entropy loss in Eq. 15:

$$L(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log p_\theta(y_i \mid x_{t,i}, x_{v,i}) \tag{15}$$

where N is the number of training examples. In our experiments, we use the AdamW optimiser with a learning rate of 2×10{-5} and weight decay of 0.01, following commonly used settings in transformer-based FND and related classification tasks [7], [8]. We train for up to 10 epochs with early stopping (patience=2) based on validation performance, and report accuracy, balanced accuracy, F1-score, AUROC, and AUPRC as our main evaluation metrics, in line with prior FND work [3], [6], [9].

For reproducibility and robustness assessment, we run experiments with multiple random seeds (42, 43, 44) and report mean and standard deviation across runs to quantify run-to-run variability.

Overview of the proposed architecture with tensor dimensions. RoBERTa and ViT produce pooled embeddings hvR768, which are projected to a shared hidden space dh=256 before a single-step cross-attention interaction. The fused representation hcmR256 is concatenated with the original unimodal embeddings to form zR1792 , which is then classified by an MLP. Because attention is applied only to pooled vectors, the figure represents coarse interaction rather than token-level or patch-level grounding.

4 Experimental Setup

4.1 Dataset

We utilise the Fakeddit dataset [30], a large-scale multimodal dataset sourced from Reddit. We focus on the 2-way classification task (True vs. Fake) using title–image pairs.

Published on June 19, 2026

Table 1: Experimental Data Settings

Setting	Train	Validation	Test
Preliminary	5000	1000	1000
Primary	50000	10000	10000

Our primary evaluation uses a larger subset with 50,000/10,000/10,000 train/validation/test samples. We also retain a smaller 5,000/1,000/1,000 setting as a preliminary experiment for faster iteration. For each random seed (42, 43, 44), we draw stratified samples without replacement from the official train, validation, and test splits while preserving the class ratio within each split. The same sampled subset is then used to compare all model variants under that seed.

Stratification is performed with respect to the binary label (True vs. Fake) within each split. For reproducibility and fair paired comparisons, we fix the sampled indices for each seed and reuse them across all model variants; the seed thus defines both the sampled subset and the stochastic training initialisation.

4.2 Implementation Details

The model was implemented in PyTorch using the Hugging Face Transformers library for encoder backbones and TIMM for ViT models. Experiments were conducted in a Kaggle GPU environment (2 × Tesla T4). The specific hyperparameters used for training are detailed in Table 2.

Table 2: Hyperparameter Configuration

Parameter	Value
Optimizer	AdamW
Learning Rate	2×10^{-5}
Batch Size	32
Epochs	10 (max), early stopping patience = 2
Dropout Rate	0.3
Hidden Dimension (d_h)	256
Weight Decay	0.01
Random Seeds	42, 43, 44

Unless otherwise stated, we keep the optimisation and regularisation hyperparameters fixed across all model variants to isolate the effect of the fusion mechanism. We use standard cross-entropy loss without class reweighting or resampling; stratification preserves the binary class proportions within each split.

4.3 Evaluation Metrics

To provide a comprehensive assessment of model performance, we report the following metrics:

- Accuracy:

  Overall correct predictions divided by total predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

  In Eq. 16 TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

- Balanced Accuracy: Average of per-class recall, useful for slightly imbalanced datasets.

$$\text{Bal Acc} = \frac{1}{2}(\text{Rec}_0 + \text{Rec}_1) \tag{17}$$

- Macro F1-Score: Unweighted mean F1-score across classes, emphasising minority class performance.

Published on June 19, 2026

$$F1_{\text{macro}} = \frac{1}{2}(F1_0 + F1_1) \tag{18}$$

- AUROC:

  Area under the receiver operating characteristic curve (AUROC), reflecting discrimination ability across decision thresholds.

- AUPRC:

  Area under the precision-recall curve (AUPRC), especially informative for imbalanced data.

All metrics are computed on the test set and reported with mean and standard deviation across three random seeds. For a small number of key model comparisons, we also report paired seed-level comparisons (paired t-test and Wilcoxon signed-rank test) as descriptive checks. Because n=3 has low statistical power, we treat these tests as descriptive and avoid strong claims of statistical superiority when differences are small.

5 Results

We evaluated five architectural configurations to separate the contribution of each modality from the effect of the fusion strategy: a text-only model, an image-only model, a simple concatenation baseline (late fusion), the proposed pooled attention fusion architecture, and a gated-fusion variant that learns a scalar modality weight.

5.1 Primary Results on 50k Samples

Table 3 presents the primary results on the larger Fakeddit subset. Three patterns stand out. First, all multimodal models outperform the unimodal baselines. Second, no single fusion strategy dominates every metric: late fusion attains the best mean accuracy and macro-F1, while gated fusion yields the best AUROC and AUPRC. Third, pooled attention fusion remains competitive, but it does not provide a consistent advantage over simpler alternatives.

5.2 Secondary Results on 5k Samples

Table 4 reports the smaller 5,000/1,000/1,000 experiment. The overall pattern is similar to the larger setting. Pooled attention fusion improves over the text-only baseline and attains the highest mean AUROC, but late fusion and especially gated fusion achieve better accuracy and macro-F1.

Table 3: Primary Results on English Fakeddit Dataset

Model	Accuracy	Macro F1	AUROC	AUPRC
Text-Only (RoBERTa)	86.08% ±0.43	85.64% ±0.31	0.9293 ±0.0027	0.8973 ±0.0037
Image-Only (ViT)	80.95% ±0.30	80.15% ±0.60	0.8476 ±0.0165	0.7776 ±0.0135
Late Fusion (Concat)	88.78% ±0.34	88.29% ±0.30	0.9505 ±0.0029	0.9172 ±0.0069
Pooled Attention Fusion (Proposed)	88.50% ±0.40	88.07% ±0.42	0.9460 ±0.0066	0.9120 ±0.0056
Gated Fusion	88.47% ±0.27	88.07% ±0.28	0.9516 ±0.0018	0.9182 ±0.0022

Table 4: Performance Comparisons of Architectures on English Fakeddit Dataset

Model	Accuracy	Bal. Acc.	Macro F1	AUROC	AUPRC
Text-Only (RoBERTa)	84.07% ±1.44	84.42% ±0.56	83.80% ±1.24	0.9182 ±0.0044	0.8923 ±0.0090
Image-Only (ViT)	76.90% ±1.35	76.56% ±2.02	76.31% ±1.65	0.7972 ±0.0180	0.7289 ±0.0248
Late Fusion (Concat)	85.73% ±0.62	86.06% ±0.52	85.47% ±0.52	0.9340 ±0.0016	0.8939 ±0.0068
Pooled Attention Fusion (Proposed)	84.90% ±0.29	85.87% ±0.28	84.78% ±0.29	0.9393 ±0.0056	0.9055 ±0.0095
Gated Fusion	87.30% ±0.99	87.03% ±1.03	86.93% ±1.00	0.9379 ±0.0036	0.9028 ±0.0061

5.3 What the Results Support

Taken together, Tables 3 and 4 support a narrower conclusion than the original draft. Multimodal fusion is clearly beneficial relative to unimodal baselines, but pooled attention fusion is not the strongest overall method in this study. Across both 5k and 50k settings, the AUROC gap between pooled attention fusion and gated fusion is small (about 0.002 in mean) and does not show a consistent advantage across seeds. With only three seeds, any seed-level paired comparisons have low power, so we avoid strong claims of statistical superiority when the observed differences between

Published on June 19, 2026

strong fusion baselines are small.

The clearest practical takeaway is that simple fusion methods are hard to beat. Gated fusion performs especially well, which suggests that in this setting learning how much to trust each modality may matter more than adding a single pooled interaction term. This interpretation is also consistent with the model design: because the pooled attention block operates on one text vector and one image vector, it cannot test fine-grained word-to-region contradictions.

5.4 Stability and Reproducibility

We report mean and standard deviation across three random seeds for all models to quantify run-to-run variability. Because the seed count is small, these variability estimates should be interpreted cautiously, and small differences between strong fusion baselines are best treated as suggestive rather than definitive.

5.5 Qualitative Error Analysis

To keep this manuscript focused on controlled quantitative evidence, we do not include qualitative case studies in the current version. A predefined disagreement-focused qualitative protocol is described in Future Work for follow-up analysis.

6 Discussion

Multimodal fusion consistently improves over unimodal baselines on Fakeddit, but no fusion strategy dominates across metrics. Accordingly, the main contribution of the present model is modest and specific rather than sweeping. These findings directly answer our central question: pooled attention provides a stable coarse interaction term, but it does not deliver consistent gains over strong simple fusion baselines in this setting.

6.1 What the Comparison Suggests

Text is the strongest unimodal source of information in our setup, and every multimodal variant improves on the image-only baseline by a large margin. The more informative comparison is therefore among fusion strategies. The clearest lesson is that simple baselines remain very strong. Late fusion is best in mean accuracy at 50k scale, and gated fusion is consistently strong across both scales. One plausible interpretation is that, for this task, learning which modality to trust is often more important than modelling a richer interaction between them. In a weakly labelled title–image setting, a flexible gating mechanism may be enough to downweight less informative visual evidence while preserving the strong textual signal.

Pooled attention fusion is still useful as a compact interaction module, but its benefits are modest. In the current implementation it adds a pooled-level compatibility term between global text and image summaries, not an explicit fine-grained inconsistency detector. That distinction helps explain why pooled attention remains competitive without emerging as the strongest model overall.

6.2 Limitations of the Current Fusion Design

The main architectural limitation is that attention is applied to pooled vectors only. This means the model cannot identify which words conflict with which image regions, nor can it offer meaningful spatial interpretability. Any interpretability claim must therefore remain limited. A stronger version of this idea would likely require token-level text features, patch-level image features, or both, which would better support fine-grained inconsistency analysis.

The empirical study also has limits. Although the 50k setting is substantially stronger than the original small-scale experiment, it is still a subset of Fakeddit, and the labels are weak. We report paired comparisons for the key 50k AUROC gap (pooled attention fusion vs. gated fusion); however, with only three seeds, statistical power is limited; we therefore treat these tests as descriptive rather than definitive. Broader inferential coverage across model pairs/metrics with multiple-comparison control and comparisons against stronger pretrained vision–language baselines are still limited. These omissions do not invalidate the current comparison, but they do bound the strength of the conclusions.

6.3 Dataset Characteristics and Generalisation

Fakeddit titles are short, Reddit-specific, and weakly labelled; images range from authentic photographs to memes and screenshots. These properties likely explain why text-only performance is already strong and why simple multimodal fusion performs so well. They also caution against broad claims of generalisation. Additional evaluation on other multimodal misinformation datasets would be needed before drawing platform-independent conclusions, even if the current trends are encouraging. This is consistent with the strong text-only baseline and the effectiveness of simple fusion observed in Tables 3 and 4.

Published on June 19, 2026

7 Conclusion

We investigated whether a single-block pooled attention fusion module improves multimodal fake news detection beyond simpler fusion strategies. On Fakeddit, multimodal models consistently outperform unimodal baselines, confirming that combining title and image information is useful. However, the primary 50k-sample evaluation shows no clear overall winner for pooled attention fusion: late fusion achieves the best mean accuracy, gated fusion achieves the best AUROC/AUPRC, and the proposed model remains competitive without consistently surpassing those baselines.

Overall, the experiments indicate that pooled attention is competitive, but it does not provide consistent gains over late fusion or gated fusion under the controlled multi-seed protocol used here.

In summary, the main takeaway is narrower but more defensible: pooled attention fusion is a stable and reasonable coarse interaction module, yet it should not be presented as evidence of fine-grained inconsistency reasoning.

7.1 Future Work

Future work will focus on four direct extensions of the current study:

- Patch-level or token-level fusion. Replace pooled-embedding attention with patch-aware image features and, where feasible, token-level text features so the model can actually test localised claim–image mismatches.
- Stronger evaluation protocol. Strengthen empirical validation by comparing against pretrained vision–language baselines and reporting performance trends across training-set size; broader pairwise comparisons across fusion variants (with appropriate multiple-comparison control) are left for future work.
- Structured qualitative error analysis. Apply a predefined disagreement-focused protocol (for example, pooled attention fusion vs. gated fusion disagreements) and report category frequencies with representative examples to improve interpretability without relying on anecdotal cases.
- Broader dataset coverage. Evaluate on larger portions of Fakeddit and on additional multimodal misinformation datasets to test cross-domain generalisation.

7.2 Limitations

This work has several limitations that should be acknowledged:

- Pooled-embedding attention only: The proposed fusion block operates on one global text vector and one global image vector, so it cannot model fine-grained word-to-region inconsistencies.
- Subset-based evaluation: Even the primary experiment uses a subset of Fakeddit rather than the full corpus, and the labels remain weak and potentially noisy.
- Limited statistical evidence: With only three random seeds, statistical power is low; we therefore interpret small differences between strong fusion baselines cautiously and do not attempt exhaustive pairwise testing across all models/metrics.
- Baseline scope: The comparison includes strong simple baselines, but not stronger pretrained vision–language models such as CLIP-style or multimodal transformer baselines [31].
- Platform specificity: The evaluation is confined to English Reddit title–image pairs, so transfer to other platforms or richer multimodal settings remains untested [32].

Despite these limitations, the paper provides a reproducible baseline and a clearer empirical message about what pooled attention fusion can and cannot currently support in multimodal fake news detection.

References

[1] M. Tajrian, A. Rahman, M. A. Kabir, and Md. R. Islam, "A Review of Methodologies for Fake News Analysis," IEEE Access, vol. 11, pp. 73879–73893, 2023, https://doi.org/10.1109/ACCESS.2023.3294989.

[2] Malliga Subramanian, Premjith B, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Balasubramanian Palani, Bharathi Raja Chakravarthi. Overview of the Shared Task on Fake News Detection in Dravidian Languages-DravidianLangTech@NAACL 2025. Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages. 2025. 759-767. https://doi.org/10.18653/v1/2025.dravidianlangtech-1.128

[3] M. Nasser et al., "A systematic review of multimodal fake news detection on social media using deep learning models," Results Eng., vol. 26, p. 104752, Jun. 2025, https://doi.org/10.1016/j.rineng.2025.104752.

[4] D. Mouratidis, A. Kanavos, and K. Kermanidis, "From Misinformation to Insight: Machine Learning Strategies for Fake News Detection," Information, vol. 16, no. 3, p. 189, Feb. 2025, https://doi.org/10.3390/info16030189.

[5] Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui, Zhoujun Li, Philip S. Yu. Kuntur, S., Wróblewska, A., Paprzycki, M., & Ganzha, M. (2025). Under the Influence: A Survey of Large Language Models in Fake News Detection. IEEE Transactions on Artificial Intelligence, 6(02), 458–476. arXiv. 2018. https://doi.org/10.1109/TAI.2024.3471735 https://arxiv.org/abs/1806.00749

[6] J. Lv, Y. Gao, L. Li, L. Shi, and S. Li, "Multi-modal fake news detection: A comprehensive survey on deep learning technology, advances, and challenges," J. King Saud Univ. Comput. Inf. Sci., vol. 37, no. 9, p. 306, 2025, https://doi.org/10.1007/s44443-025-00317-7.

[7] Y. Chen et al., "Cross-modal Ambiguity Learning for Multimodal Fake News Detection," in Proceedings of the ACM Web Conference 2022, Virtual Event, Lyon France: ACM, Apr. 2022, pp. 2897–2905. https://doi.org/10.1145/3485447.3511968.

[8] L. Peng, S. Jian, Z. Kan, L. Qiao, and D. Li, "Not all fake news is semantically similar: Contextual semantic representation learning for multimodal fake news detection," Inf. Process. Manag., vol. 61, no. 1, p. 103564, Jan. 2024, https://doi.org/10.1016/j.ipm.2023.103564.

[9] X. Fang et al., "NSEP: Early fake news detection via news semantic environment perception," Inf. Process. Manag., vol. 61, no. 2, p. 103594, Mar. 2024, https://doi.org/10.1016/j.ipm.2023.103594.

[10] J. Jing, H. Wu, J. Sun, X. Fang, and H. Zhang, "Multimodal fake news detection via progressive fusion networks," Inf. Process. Manag., vol. 60, no. 1, p. 103120, Jan. 2023, https://doi.org/10.1016/j.ipm.2022.103120.

[11] L. Hu, S. Wei, Z. Zhao, and B. Wu, "Deep learning for fake news detection: A comprehensive survey," AI Open, vol. 3, pp. 133–155, 2022, https://doi.org/10.1016/j.aiopen.2022.09.001.

[12] H. T. Phan, N. T. Nguyen, and D. Hwang, "Fake news detection: A survey of graph neural network methods," Appl. Soft Comput., vol. 139, p. 110235, May 2023, https://doi.org/10.1016/j.asoc.2023.110235.

[13] A. K. Gupta and M. P. Singh, "Self and cross-modal attention based features fusion for fake news detection," Multimed. Tools Appl., vol. 85, no. 1, p. 10, Jan. 2026, https://doi.org/10.1007/s11042-026-21186-w.

[14] A. J. Dal Forno, G. P. Richetti, and V. H. Knaesel, "Fake news detection algorithms – A systematic literature review," Data Knowl. Eng., vol. 158, p. 102441, Mar. 2025, https://doi.org/10.1016/j.datak.2025.102441

[15] F. A. Alshuwaier and F. A. Alsulaiman, "Fake News Detection Using Machine Learning and Deep Learning Algorithms: A Comprehensive Review and Future Perspectives," Computers, vol. 14, no. 9, 2025, https://doi.org/10.3390/computers14090394.

[16] F. Rauf, R. Irfan, L. Mushtaq, and M. Ashraf, "Fake News Detection in Urdu using Deep Learning," VFAST Trans. Softw. Eng., vol. 10, no. 4, pp. 151–167, Dec. 2022, https://doi.org/10.21015/vtse.v10i4.1290.

[17] A. Rafique, F. Rustam, M. Narra, A. Mehmood, E. Lee, and I. Ashraf, "Comparative analysis of machine learning methods to detect fake news in an Urdu language corpus," PeerJ Comput. Sci., vol. 8, p. e1004, Jun. 2022, https://doi.org/10.7717/peerj-cs.1004.

[18] M. Wasim, S. M. Cheema, and I. M. Pires, "Normalized effect size (NES): a novel feature selection model for Urdu fake news classification," PeerJ Comput. Sci., vol. 9, p. e1612, Oct. 2023, https://doi.org/10.7717/peerj-cs.1612.

[19] E. Hashmi, S. Y. Yayilgan, M. M. Yamin, S. Ali, and M. Abomhara, "Advancing Fake News Detection: Hybrid Deep Learning With FastText and Explainable AI," IEEE Access, vol. 12, pp. 44462–44480, 2024, https://doi.org/10.1109/ACCESS.2024.3381038.

[20] Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui, Zhoujun Li, Philip S. Yu. TI-CNN: Convolutional Neural Networks for Fake News Detection. arXiv. 2018. https://arxiv.org/abs/1806.00749

[21] F. Farhangian, R. M. O. Cruz, and G. D. C. Cavalcanti, "Fake news detection: Taxonomy and comparative study," Inf. Fusion, vol. 103, p. 102140, 2024, https://doi.org/10.1016/j.inffus.2023.102140.

[22] P. He, H. Zhang, S. Cao, and Y. Wu, "Cross-Modal Fake News Detection Method Based on Multi-Level Fusion Without Evidence," Algorithms, vol. 18, no. 7, 2025, https://doi.org/10.3390/a18070426.

[23] Y. Jiang, T. Wang, X. Xu, Y. Wang, X. Song, and D. Maynard, "Cross-modal augmentation for few-shot multimodal fake news detection," Eng. Appl. Artif. Intell., vol. 142, p. 109931, 2025, https://doi.org/10.1016/j.engappai.2024.109931.

[24] B. Hu, Z. Mao, and Y. Zhang, "An overview of fake news detection: From a new perspective," Fundam. Res., vol. 5, no. 1, pp. 332–346, Jan. 2025, https://doi.org/10.1016/j.fmre.2024.01.017.

[25] K. I. Roumeliotis, N. D. Tselikas, and D. K. Nasiopoulos, "Fake News Detection and Classification: A Comparative Study of Convolutional Neural Networks, Large Language Models, and Natural Language Processing Models," Future Internet, vol. 17, no. 1, p. 28, Jan. 2025, https://doi.org/10.3390/fi17010028.

[26] L. Wu, Y. Long, C. Gao, Z. Wang, and Y. Zhang, "MFIR: Multimodal fusion and inconsistency reasoning for explainable fake news detection," Inf. Fusion, vol. 100, p. 101944, 2023, https://doi.org/10.1016/j.inffus.2023.101944.

[27] C.-O. Truică, E.-S. Apostol, M. Marogel, and A. Paschke, "GETAE: Graph Information Enhanced Deep Neural NeTwork Ensemble ArchitecturE for fake news detection," Expert Syst. Appl., vol. 275, p. 126984, May 2025, https://doi.org/10.1016/j.eswa.2025.126984.

[28] Xiaochuan Xu, Peiyang Yu, Zeqiu Xu, Jiani Wang. A Hybrid Attention Framework for Fake News Detection with Large Language Models. 2025 5th International Conference on Neural Networks, Information and Communication Engineering (NNICE). 2025. 587-590. https://doi.org/10.1109/nnice64954.2025.11064493

[29] M. I. Nadeem et al., "HyproBert: A Fake News Detection Model Based on Deep Hypercontext," Symmetry, vol. 15, no. 2, p. 296, Jan. 2023, https://doi.org/10.3390/sym15020296.

[30] Fakeddit A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. https://doi.org/10.48550/arXiv.1911.03854

[31] Learning Transferable Visual Models From Natural Language Supervision. (n.d.). Retrieved June 17, 2026, from https://proceedings.mlr.press/v139/radford21a

[32] Grace Luo, Trevor Darrell, Anna Rohrbach. NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021. 6801-6817. https://doi.org/10.18653/v1/2021.emnlp-main.545