TriMFB: A Tri-Feature Fusion Framework Integrating Modality-Specific and Cross-Modal Representations for Fine-Grained Fake News Detection

Idza Aisara Norabid[1], Masita Jalil[2], Noor Hafhizah Abd Rahim[3]*
Artificial Intelligence Group-Faculty of Computer Science and Mathematics,
Universiti Malaysia Terengganu, Kuala Terengganu, Malaysia[1, 3]
Software Engineering Group-Faculty of Computer Science and Mathematics,
Universiti Malaysia Terengganu, Kuala Terengganu, Malaysia[2]

***Abstract*—The rapid spread of multimodal misinformation on social media platforms has strengthened the need for effective fake news detection systems. While recent multimodal approaches integrate textual and visual information, most existing studies formulate the task as binary classification, thereby failing to fully exploit the rich label structures available in fine-grained datasets. Moreover, conventional fusion strategies such as simple concatenation often fail to capture complex cross-modal interactions. To address these limitations, this study proposes TriMFB, an attention-enhanced Multimodal Factorized Bilinear (MFB)-based early fusion framework for fine-grained multimodal fake news detection. Unlike other frameworks, the proposed TriMFB simultaneously integrates modality-specific features, attention-refined features, and MFB-based cross-modal interactions within a unified tri-feature early fusion architecture. The proposed model extracts textual and visual representations using Bidirectional Encoder Representations from Transformers (BERT) and a 50-layer Residual Network (ResNet50), respectively. Modality-specific features are refined through attention mechanisms and integrated using MFB pooling to capture cross-modal interactions. Experiments conducted on the Fakeddit dataset demonstrate that the proposed framework outperforms baseline bilinear fusion models, achieving a classification accuracy of 0.78 in a six-class setting. Although the overall accuracy remains comparable, the proposed tri-feature fusion strategy substantially improves the detection of challenging minority classes, increasing the F1-score for the Manipulated class from 0.21 to 0.29. These findings suggest that this fusion strategy offers a practical route to more reliable detection of subtly manipulated content, a minority class that is both harder to detect and disproportionately harmful when missed by content moderation systems.**

Keywords*—Bilinear pooling; early fusion; fake news detection; fine-grained; multimodal

I. Introduction

Fake news has become a persistent threat in today's digital landscape, influencing public opinion, shaping political perspectives, and contributing to social unrest [1]. The rapid dissemination of false information across social media platforms has intensified the need for reliable automated fake news detection systems [2]. As online content increasingly combines text with images, detection systems that fail to account for this shift risk missing critical signals of manipulation, making the problem more urgent and more difficult to solve than in the text-only era.

Traditionally, news articles have been dominated by textual content, and early research in fake news detection therefore focused primarily on text-based, or unimodal, approaches, reaching nearly 99% accuracy [3], [4]. Although these methods have achieved considerable progress, their performance has begun to plateau, particularly as online news consumption increasingly incorporates visual components. With the evolution of news media, images are now frequently used to complement textual narratives. This shift has highlighted the limitations of unimodal systems, which overlook the interdependencies between textual and visual information [5], [6]. Since text and images often function as complementary cues, ignoring one modality may risk losing essential cross-modal relationships that can signal manipulation or misinformation [7].

These developments underscore the growing importance of multimodal approaches in fake news detection. Multimodal models typically differ in how and when information from different modalities is integrated, commonly through early fusion, late fusion or hybrid strategies. Early fusion, which integrates features from different modalities at the representation level [8] has been explored by several researchers for its ability to model cross-modal interactions more effectively [9], [10], [11]. However, most existing multimodal studies still focus on binary classification, which restricts the expressive richness of multimodal datasets and results in underutilized multimodal representations. Notably, fine-grained datasets such as Fakeddit [12] contain multiple categories of misleading content, yet several studies simplify the task to a binary decision [13], [14], thereby limiting the detection system's real-world relevance.

*Corresponding author.
Financial support for this study was provided by the Fundamental Research Grant Scheme, Ministry of Higher Education Malaysia.

Therefore, this study makes three contributions. First, it proposes TriMFB, a tri-feature multimodal early fusion framework for fine-grained fake news detection. Second, the framework integrates complementary textual, visual and multimodal interaction features to enhance cross-modal representation learning. Third, the proposed approach is evaluated on the Fakeddit dataset to assess its performance in fine-grained fake news classification compared with existing multimodal fusion methods.

The remainder of this study is organized as follows. Section II reviews related studies on multimodal fusion techniques and fine-grained classification. Section III introduces the proposed attention-enhanced MFB-based early fusion framework for fine-grained multimodal fake news detection. Section IV presents the experimental setup, followed by the presentation and discussion of the experimental results. Finally, Section V concludes the study and outlines potential directions for future work.

II. Related Work

A. Overview of Multimodal Fusion

Multimodal is a combination of more than one type of modality or information from multiple sources [15]. A research problem is considered multimodal when it integrates information from several distinct data types [16]. These sources may include textual or spoken language, visual content or auditory cues. In addition, multimodal data may also encompass other forms such as numerical or biological data.

In the context of fake news detection, the rise of social media has intensified the need for multimodal analysis. Online content increasingly combines textual and visual elements to enhance engagement and credibility, leading to a shift from single modality to multimodal communication. This evolution poses new challenges in identifying misinformation as deceptive content often relies on inconsistencies between textual and visual information. Consequently, multimodal fusion has become essential for integrating heterogeneous information sources such as text and images to improve detection performance. The effectiveness of multimodal learning largely depends on the fusion stage, which determines how and when information from different modalities is combined.

Fusion can generally be categorized into early fusion, late fusion, and hybrid fusion [16], [17]. Each approach has its own strengths and trade-offs in capturing cross-modal correlations and model interpretability. Early fusion, also known as feature-level fusion, combines the features extracted from different modalities at an early stage to capture interactions within the data representations. In contrast, late fusion, or decision-level fusion, involves training separate models for each modality and then combining their outputs. Finally, hybrid fusion combines outputs from early fusion and individual unimodal predictors.

B. Early Fusion Techniques

Early fusion serves as the stage where information from different modalities is integrated at the feature level. Formally, let $f_t, f_i, \dots, f_n$ represent the features extracted from each modality. The fusion process can be expressed as:

$$x = F(f_t, f_i, \dots, f_n) \quad (1)$$

where, $F(\cdot)$ denotes the fusion function that combines these multimodal features into a unified vector representation. This fused feature vector, x, is then passed to a classifier $f(\cdot)$ to produce the final prediction output y:

$$y = f(x) \quad (2)$$

Once the textual and visual features are extracted, various fusion techniques can be applied to determine how these representations are combined. The simplest and most widely used operation is concatenation, which directly merges feature vectors from each modality into a unified representation. Several studies have adopted concatenation for early fusion [18], [19]. However, this operation merely stacks features and often fails to capture the complex relationships between textual and visual data [20].

Consequently, more expressive fusion strategies have been explored. Bilinear pooling offers an alternative to conventional fusion methods. Rather than merging features directly, it models the multiplicative relationships between them by computing the outer product of the text and image feature vectors. This operation enables every dimension of the textual representation to interact with every dimension of the visual representation [21]. However, the full outer product is computationally expensive [16], motivating the development of more efficient variants.

Practical variants such as Multimodal Compact Bilinear Pooling (MCB) [22] and Multimodal Factorized Bilinear Pooling (MFB) [23] were introduced to compress or factorize the bilinear interactions, significantly reducing computational cost while preserving the expressive benefits of bilinear fusion, making them well-suited for deep learning applications. While MCB and MFB were initially developed for the visual question answering domain, they have since been adapted to fake news detection. [24] and [25] applied MCB as their fusion method, though the limitation of MCB requiring high-dimensional output features has been documented [23]. Meanwhile, [26] adopted MFB for its reduced dimensionality and computational efficiency.

In addition to advances in fusion strategies, researchers have increasingly explored attention mechanisms introduced by [27]. It enables models to dynamically assign weights to the most informative features, whether within a modality or across modalities. Particularly, self-attention enables each element within a representation to attend to other elements in the same sequence, capturing contextual dependencies through learned similarity weights. Multi-head attention extends this mechanism by projecting features into multiple subspaces, enabling the model to learn diverse interaction patterns simultaneously. In the context of fake news detection, attention has been shown to highlight relevant textual cues, emphasize important visual regions and strengthen cross-modal alignments that signal inconsistencies between modalities [28], [29].

C. Fine-Grained Setting

Most existing multimodal studies continue to frame fake news detection as a binary classification problem, despite the availability of datasets that support more fine-grained categorization such as Fakeddit. For example, a few works [13], [14], [30] employed the Fakeddit dataset but reduced the task to

binary classification, thereby underutilizing its fine-grained labels and limiting the exploitation of multimodal relationships.

Although binary classification simplifies the learning objective, it does not reflect the real-world complexity of misinformation. Fine-grained datasets often exhibit class imbalance, where certain misinformation categories contain significantly fewer samples, potentially biasing the learning process toward majority classes [31], [32].

Some researchers have explored fine grained classification.[33], [34] adopted a fine-grained setting but relied on simple concatenation for fusion. As previously discussed, concatenation merely stacks modality-specific features and may overlook complex intermodal dependencies. Meanwhile, studies utilizing more expressive fusion techniques, such as MCB or MFB, remain confined to binary classification settings [24], [25], [26]. As a result, the potential of expressive bilinear fusion has not yet been fully explored in fine grained multimodal fake news detection.

Motivated by these findings, this work proposes an early fusion framework that employs the MFB method enhanced with an attention mechanism applied to a fine-grained classification task. Fine-grained classification enables a more detailed understanding of misinformation categories, while attention-enhanced bilinear fusion strengthens cross-modal interactions, addressing the limitations observed in prior binary and structurally simplistic fusion approaches.

III. METHODOLOGY

This section outlines the proposed early fusion framework, which aims to integrate textual and visual information for fine-grained fake news classification. The proposed framework as shown in Fig. 1, comprises three main components that work in a unified manner to achieve effective multimodal representation learning. Algorithm 1 summarizes the main steps involved in the proposed framework to provide a clearer view of the workflow.

Algorithm 1: Attention-enhanced MFB-based early fusion framework

Initialize Input: Text-image pair (T, I)
Output: Predicted class label y
Initialize:
BERT
ResNet50
Attention mechanism
MFB pooling module
MLP classifier
Compute:
$f_t \leftarrow$ BERT(T)
$f_i \leftarrow$ ResNet50(I)
Refine Features:
$\hat{f}_t \leftarrow$ Attention(f_t)
$\hat{f}_i \leftarrow$ Attention(f_i)
Fuse Features:
$f_{mfb} \leftarrow$ MFB(f_i, f_i)
Construct Tri-Feature Representation:
$F \leftarrow$ concat(f_t, f_i, $\hat{f}_t$, $\hat{f}_i$, f_{mfb})
Classify:
$y \leftarrow$ MLP(F)
Return y

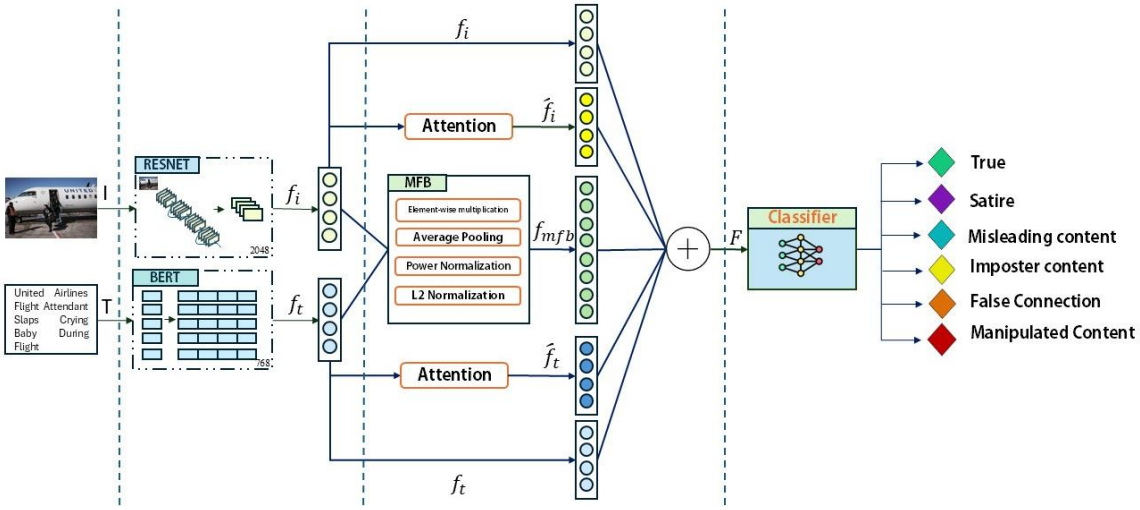

Fig. 1. Overview of the early fusion architecture with MFB pooling and attention mechanism.

The first component focuses on modality-specific feature extraction using pretrained deep neural networks, BERT [35] and ResNet50 [36]. The second component involves feature refinement and fusion, where attention mechanisms [27] are applied to capture intra-modal dependencies within each modality, while the MFB pooling [23] module is employed to learn inter-modal relationships between text and image features. The third component performs classification based on a concatenated tri-feature representation that combines the modality-specific features, attention-refined features, and the fused information into a comprehensive multimodal vector. Inspired by [25], the proposed framework adopts MFB in place of MCB and extends the architecture for fine-grained fake news classification.

The framework begins by collecting text–image pairs from the Fakeddit dataset [12]. The dataset contains approximately

one million Reddit posts labelled into six fine-grained categories, namely, true, satire, misleading content, imposter content, false connection, and manipulated content. This dataset provides a richer labelling structure compared to traditional binary classification datasets.

To begin with, the textual information from the dataset is processed using the BERT model to obtain rich contextual embeddings. Each input sentence $T = \{t_1, t_2, \ldots, t_n\}$ is first tokenized and then passed through BERT's layers to generate deep semantic representations. This embedding serves as the textual feature vector $f_t \in \mathbb{R}^{768}$. For the visual modality, each news image is passed through a pretrained ResNet model, specifically ResNet-50 in this study, to obtain deep visual representations. The output is extracted as the image feature vector $f_i \in \mathbb{R}^{2048}$. These extracted visual features capture high-level semantic information helping the model identify potential inconsistencies between textual and visual content in fake news posts.

The combination of BERT and ResNet has been widely adopted in prior multimodal studies due to their complementary strengths in capturing linguistic and visual semantics. For example, [25], [37], [38] employed pretrained BERT and ResNet models to generate high-quality representations of text and images. Motivated by these findings, this study adopts the BERT and ResNet50 combination to ensure both modalities contribute rich representations for multimodal fake news detection.

Moving to the next component, the MFB pooling method is employed to capture the complex interactions between textual and visual features. Unlike simple feature merging operations, MFB models second-order multiplicative relationships between modalities that enable richer cross-modal representations [26]. Given the textual feature$f_t \in \mathbb{R}^{768}$ and the image feature $f_i \in \mathbb{R}^{2048}$, the MFB module computes their joint representation as:

$$f_{mfb} = MFB(f_t, f_i) \tag{3}$$

Here, f_{mfb} denotes the fused multimodal feature. This fused representation integrates complementary semantic cues from both text and image modalities, bridging linguistic meaning and visual context.

In the proposed framework, the textual and visual features have dimensions of 768 and 2048, respectively. Each feature is first projected into a factorized space using separate linear layers. The MFB output dimension is set to 512 with five factors, resulting in a projection dimension of 2560. The projected features are combined through element-wise multiplication, followed by dropout with a rate of 0.1. The resulting representation is reshaped to 512 x 5 and sum-pooled across the five factors to obtain the 512-dimensional fused feature. Finally, signed square-root transformation and L2 normalization are applied to the pooled representation.

In parallel with the MFB fusion process, the modality-specific features are refined using attention mechanisms. Self-attention and multi-head attention are separately applied to the textual and visual features to emphasize informative components within each modality. The refined textual and image features are represented as $\widehat{f_t} = Attn(f_t)$ and $\widehat{f_i} = Attn(f_i)$, respectively. This refinement step allows each modality to emphasize relevant information before the final multimodal representation is constructed.

Finally, the tri-feature representation, the modality-specific features of both textual and visual, f_t *and* f_i, the attention-refined features, ($\widehat{f_t}$) and ($\widehat{f_i}$), and fused (f_{mfb}) are concatenated to construct a unified multimodal feature vector, expressed as:

$$F = concat[f_t, f_i, \widehat{f_t}, \widehat{f_i}, f_{mfb,}\ ] \tag{4}$$

where, $f_t \in \mathbb{R}^{768}$, $f_i \in \mathbb{R}^{2048}$, $\widehat{f_t} \in \mathbb{R}^{768}$ and $\widehat{f_i} \in \mathbb{R}^{2048}$. Thus, the resulting tri-feature representation has a dimension of 6144. The attention refinement and MFB fusion are performed independently, with both operations receiving the modality-specific textual and visual features as inputs. By combining these complementary representations, the model achieves a more comprehensive understanding of multimodal content. The concatenated feature is then fed to the MLP classifier to produce the final predicted news label for fine-grained classification.

IV. Experiment and Discussion

A. Dataset

The experiments were conducted using the Fakeddit dataset [12], which serves as the benchmark dataset for fine-grained multimodal fake news detection. Fakeddit contains approximately one million Reddit posts labelled with six categories, which are true, satire, misleading content, imposter content, false connection, and manipulated content. Each sample contained a textual component and its corresponding image, enabling a multimodal analysis. A subset of 20,000 text-image pairs was used to balance computational efficiency and predictive performance, as previous findings [39] indicate that performance gains beyond this size are marginal. This experimental scale is also consistent with prior fake news detection studies that employed datasets of similar size [40], [41].

B. Experiment Setup

All experiments were implemented in Python using the PyTorch deep learning framework. The models were trained on a machine equipped with an Intel Core i7 processor and 16 GB of RAM. To simulate real-world resource limitations, all training and evaluation were conducted in a CPU-based environment. The Fakeddit dataset served as the primary benchmark, supporting fine-grained fake news classification with six distinct labels. The dataset was divided into 80% training and 20% testing. Each model was trained independently using the same data splits. Cross-validation and statistical significance testing were not performed in this study. The class distribution of the selected subset is presented in Table I. This imbalance may affect classification performance, particularly for minority categories, and is considered when interpreting the per-class metrics.

The proposed framework employs BERT to extract textual features and ResNet50 to extract visual features. The MFB fusion strategy was employed in an early fusion setting, where textual and visual features were combined before classification. Each modality-specific feature was further refined through attention layers to emphasize informative components within

each modality. The tri-feature representation is then concatenated before being passed to an MLP classifier for final prediction.

TABLE I. CLASS DISTRIBUTION OF THE 20,000-SAMPLE FAKEDDIT SUBSET

Class	Training Set	Validation Set	Total
True	6,369	1,572	7,941
Satire	976	239	1,215
Misleading	3,004	801	3,805
Manipulated	337	72	409
False Connection	4,716	1,165	5,881
Imposter Content	598	151	749
Total	16,000	4,000	20,000

C. Hyperparameter Setting

Training was conducted for up to 10 epochs using the Adam optimizer with a learning rate of 1e-4 and a batch size of 32. The pretrained BERT and ResNet-50 parameters were fine-tuned during training. Early stopping was applied to prevent overfitting. Table II summarizes the training configuration settings employed in this work.

TABLE II. TRAINING CONFIGURATION SETTINGS

Parameter	Value
Optimizer	Adam
Learning Rate	1×10^{-4}
Batch Size	32
Number of Epochs	10
Loss Function	CrossEntropyLoss
Train-Test Split	80:20
Random Seed	42
Regularization	Early stopping + dropout
Dropout Rate	0.3

D. Evaluation Metrics

Model performance was evaluated using accuracy, precision, recall, and F1-score. Accuracy was used to measure the overall classification performance. Precision, recall, and F1-score provided a more detailed evaluation of the model's performance across the six fine-grained classes.

E. Result and Discussion

This section discusses the experimental results obtained from the proposed early fusion multimodal fake news detection framework. The evaluation focuses on analysing the impact of different attention mechanisms on refining the textual and visual features before fusion. In the proposed model, self-attention and multi-head attention are applied separately to enhance the discriminative quality of each modality. Through comparative analysis, this section aims to highlight the strengths and limitations of each configuration, providing insights into how attention mechanisms contribute to improving multimodal representation learning and classification performance.

Table III presents the comparative performance of the proposed framework against baseline models. The baseline implementations by [26], [24], and [25] achieved accuracies of 0.59, 0.70, and 0.75, respectively. In contrast, all variants of the proposed model consistently outperformed these baselines, achieving accuracies ranging from 0.77 to 0.78. This improvement suggests that integrating attention mechanisms, as introduced by [24], together with MFB pooling, can provide useful multimodal representations for fine-grained fake news classification.

TABLE III. ACCURACY COMPARISON BETWEEN BASELINE MODELS

Baseline model	Modality-specific features	Attention mechanism	Accuracy
AMFB [26]	No	Additive attention	0.59
Double-MCBP [24]	No	Self-attention	0.75
MBPAM [25]	Yes	Self-attention	0.70
Proposed framework	Yes	Self-attention	0.78

F. Ablation Study on Attention Mechanisms

When examining the impact of attention mechanisms, the results show minimal variation across all settings in Table IV, with accuracies ranging from 0.77 to 0.78. Without modality-specific feature inclusion, both self-attention and multi-head attention achieved an identical accuracy of 0.78. With modality-specific features included, self-attention maintained an accuracy of 0.78, while multi-head attention showed a slight decrease to 0.77. These findings indicate that the choice between self-attention and multi-head attention has little impact on the overall classification performance. Nevertheless, a closer examination of the per-class metrics reveals a more comprehensive assessment across categories.

TABLE IV. ACCURACY COMPARISON BETWEEN ATTENTION MECHANISM

Modality-specific features	Attention mechanism	Accuracy
No	Self-attention	0.78
No	Multi-head Attention	0.78
Yes	Self-attention (Proposed)	0.78
Yes	Multi-head Attention	0.77

TABLE V. PER-CLASS METRICS OF THE SELF-ATTENTION WITHOUT MODALITY-SPECIFIC FEATURE INTEGRATION

Class	Precision	Recall	F1-Score
True	0.82	0.81	0.81
Satire	0.64	0.53	0.58
Misleading	0.70	0.67	0.69
Manipulated	0.21	0.21	0.21
False Connection	0.86	0.92	0.89
Imposter Content	0.62	0.63	0.63

Tables V–VIII present the per-class metrics of the proposed attention-based MFB framework under different configurations, examining the effects of attention mechanisms and the inclusion of modality-specific features. Overall, the results indicate that the model consistently performs well on majority classes, particularly True and False Connection, which achieve high F1-scores across all configurations. This suggests that the fusion

strategy effectively captures dominant multimodal patterns in well-represented categories.

TABLE VI. PER-CLASS METRICS OF THE MULTI-HEAD ATTENTION WITHOUT MODALITY-SPECIFIC FEATURE INTEGRATION

Class	Precision	Recall	F1-Score
True	0.80	0.83	0.82
Satire	0.67	0.44	0.53
Misleading	0.75	0.62	0.68
Manipulated	0.20	0.36	0.25
False Connection	0.86	0.93	0.90
Imposter Content	0.62	0.68	0.65

When modality-specific features are excluded (Tables V and VI), both self-attention and multi-head attention show comparable performance, with slight variations across minority classes. Multi-head attention shows marginal improvements in certain categories such as Manipulated and Imposter Content, with F1-scores increasing from 0.21 to 0.25 and from 0.63 to 0.65, respectively. However, the differences remain relatively small, suggesting that the choice of attention variant provides comparable benefits for feature refinement.

TABLE VII. PER-CLASS METRICS OF THE SELF-ATTENTION WITH MODALITY-SPECIFIC FEATURE INTEGRATION (PROPOSED)

Class	Precision	Recall	F1-Score
True	0.75	0.89	0.81
Satire	0.85	0.34	0.49
Misleading	0.69	0.67	0.68
Manipulated	0.38	0.24	0.29
False Connection	0.92	0.84	0.87
Imposter Content	0.75	0.72	0.73

TABLE VIII. PER-CLASS METRICS OF THE MULTI-HEAD ATTENTION WITH MODALITY-SPECIFIC FEATURE INTEGRATION

Class	Precision	Recall	F1-Score
True	0.78	0.85	0.81
Satire	0.75	0.47	0.58
Misleading	0.68	0.63	0.65
Manipulated	0.22	0.19	0.20
False Connection	0.87	0.89	0.88
Imposter Content	0.70	0.57	0.63

With modality-specific features included (Tables VII and VIII), performance improvements are observed in selected classes. In particular, self-attention achieves a higher F1-score for Manipulated at 0.29 and Imposter Content at 0.73, while multi-head attention achieves a slightly higher F1-score for False Connection at 0.88. These results suggest that incorporating modality-specific features alongside attention-refined and fused representations may provide complementary information for distinguishing challenging categories. However, the improvements are not consistent across all classes, indicating that the effect of modality-specific feature integration varies across attention configurations.

Across all configurations, Manipulated consistently exhibits lower F1-scores, which may be related to class imbalance and higher semantic ambiguity. As shown in Table I, the dataset is highly imbalanced across the six classes, with Manipulated representing the smallest category with 409 samples. This imbalance may contribute to the lower performance observed for minority categories. This observation reinforces the challenge highlighted in the literature regarding fine-grained fake news classification, where minority and subtle misinformation categories are inherently more difficult to distinguish than majority classes. Despite this difficulty, the proposed framework maintains stable performance across configurations, suggesting its ability to handle complex class distinctions.

The inclusion of modality-specific features contributes to improved performance in certain underrepresented categories, particularly under the self-attention configuration. This suggests that combining modality-specific features with attention-refined and fused representations provides complementary discriminative information. This finding highlights the value of richer multimodal representations for distinguishing fine-grained misinformation categories, as discussed in previous research.

The results are consistent with the motivation outlined in Section II. By employing attention-enhanced MFB fusion within a fine-grained classification setting, the proposed framework provides a richer representation of cross-modal information. Compared to prior binary-focused approaches, the tri-feature design enables more fine-grained representation learning, which is reflected in the model's competitive and stable performance across diverse misinformation categories.

G. Ablation Study

Table IX compares the performance of different feature configurations within the proposed framework. Four configurations, namely Modality-Specific Features, Modality-Specific + Attention-Refined Features, Attention-Refined + MFB Fused Features, and the proposed framework, all achieved the highest overall accuracy of 0.78. Since these configurations exhibit identical overall accuracy, a more detailed analysis using per-class precision, recall, and F1-score is conducted to better understand their classification behaviour across individual misinformation categories.

Although all four configurations achieved the same overall accuracy of 0.78, a closer examination of the per-class metrics presented in Tables V, VII, X and XI reveals differences in classification performance across individual categories. This similarity in overall accuracy may be influenced by the dominance of majority classes, particularly True and False Connection, for which the model achieves relatively high performance. Therefore, overall accuracy alone may not fully reflect differences in the classification of minority categories.

In particular, the Manipulated class, which contains the fewest training samples, shows consistently lower performance across the configurations. As shown in Tables X and XI, incorporating attention-refined features alongside modality-specific features improves the F1-score for Manipulated from 0.22 to 0.26. Similarly, comparing the attention-enhanced MFB configuration in Table V with the proposed TriMFB framework in Table VII shows a further improvement from 0.21 to 0.29.

These results suggest that integrating modality-specific, attention-refined, and MFB-fused representations provides complementary information that may help the model distinguish challenging minority categories.

TABLE IX. ACCURACY COMPARISON BETWEEN CONFIGURATIONS

Configurations	Feature representation	Fusion technique	Accuracy
Modality-specific features	1. Raw textual 2. Raw visual	Concatenation	0.78
Attention-refined features	1. Attention-refined textual 2. Attention-refined visual	Concatenation	0.77
MFB fused features	1. Raw textual 2. Raw visual	MFB	0.68
Attention-refined MFB fusion features	1. Attention-refined textual 2. Attention-refined visual	MFB	0.71
Modality-specific + Attention-refined features	1. Raw textual 2. Raw visual 3. Attention-refined textual 4. Attention-refined visual	Concatenation	0.78
Modality-specific + MFB fused features	1. Raw textual 2. Raw visual 3. MFB fused	MFB + concatenation	0.76
Attention-refined + MFB fused features	1. Attention-refined textual 2. Attention-refined visual 3. MFB fused	MFB + concatenation	0.78
Modality-specific + Attention-refined features + MFB fused features (Proposed)	1. Raw textual 2. Raw visual 3. Attention-refined textual 4. Attention-refined visual 5. MFB fused	MFB + concatenation	0.78

TABLE X. PER-CLASS METRIC FOR MODALITY-SPECIFIC FEATURES

Class	Precision	Recall	F1-Score
True	0.75	0.89	0.81
Satire	0.67	0.43	0.52
Misleading	0.74	0.59	0.66
Manipulated	0.26	0.19	0.22
False Connection	0.90	0.90	0.90
Imposter Content	0.61	0.54	0.57

TABLE XI. PER-CLASS METRIC FOR MODALITY-SPECIFIC + ATTENTION-REFINED FEATURES

Class	Precision	Recall	F1-Score
True	0.80	0.81	0.81
Satire	0.58	0.67	0.62
Misleading	0.68	0.68	0.68
Manipulated	0.29	0.24	0.26
False Connection	0.88	0.89	0.88
Imposter Content	0.78	0.50	0.61

This observation is consistent with the literature review, which highlights the need for expressive multimodal fusion strategies to capture complex intermodal relationships in fine-grained fake news detection. The gradual improvement in the Manipulated class suggests that the proposed tri-feature representation provides complementary information that is not captured by any single feature representation alone.

V. CONCLUSION

This study proposes TriMFB, an attention-enhanced MFB-based early fusion framework for fine-grained multimodal fake news detection. Unlike prior works that primarily focused on binary classification, the proposed framework leverages the full label richness of the Fakeddit dataset to better reflect the complexity of real-world misinformation. By integrating modality-specific features, attention-refined features, and MFB-fused features, the framework creates a unified tri-feature representation that captures both important features from each modality and the relationships between text and images. Experimental results show that the proposed framework achieves an accuracy of 0.78, outperforms baseline bilinear fusion models, and maintains stable performance across different attention configurations. Although performance remains lower for minority categories, which may be related to class imbalance and semantic ambiguity, the framework shows competitive behavior across fine-grained classes. These findings suggest the potential of combining attention mechanisms with bilinear fusion for fine-grained fake news classification. The proposed framework was evaluated using only the Fakeddit dataset, which may limit its generalizability to other misinformation domains and social media platforms. Future work will focus on validating the framework using additional benchmark datasets, cross-validation and statistical significance testing, while exploring advanced imbalance mitigation strategies to further improve minority class performance.

ACKNOWLEDGMENT

Financial support for this study was provided by the Fundamental Research Grant Scheme, Ministry of Higher Education Malaysia, designated by grant number FRGS/1/2023/ICT02/UMT/03/1.

REFERENCES

[1] E. Mwangi, “Technology and Fake News: Shaping Social, Political, and Economic Perspectives,” SSRN Electronic Journal, 2023, doi: 10.2139/ssrn.4462727.

[2] B. Collins, D. T. Hoang, N. T. Nguyen, and D. Hwang, “Trends in combating fake news on social media–a survey,” Journal of Information and Telecommunication, vol. 5, no. 2, pp. 247–266, 2021, doi: 10.1080/24751839.2020.1847379.

[3] R. H. Al-Furaiji and H. Abdulkader, “A Comparison of the Performance of Six Machine Learning Algorithms for Fake News,” EAI Endorsed Transactions on AI and Robotics, vol. 3, Jan. 2024, doi: 10.4108/airo.4153.

[4] E. Fayez, A. E. Aboutabl, and S. N. Abdulkader, “Automated detection of fake news,” International Journal of Informatics and Communication Technology, vol. 12, no. 1, pp. 79–84, Apr. 2023, doi: 10.11591/ijict.v12i1.pp79-84.

[5] S. Zhong, W. Ruan, M. Jin, H. Li, Q. Wen, and Y. Liang, “Time-VLM: Exploring Multimodal Vision-Language Models for Augmented Time Series Forecasting,” May 2025, [Online]. Available: http://arxiv.org/abs/2502.04395

[6] W. Chen, Y. Dang, and X. Zhang, "A Multimodal Semantic-Enhanced Attention Network for Fake News Detection," Entropy, vol. 27, no. 7, Jul. 2025, doi: 10.3390/e27070746.

[7] E. I. Setiawan et al., "An image and text-based fake news detection with transfer learning," PLoS One, vol. 20, no. 6, p. e0324394, Jun. 2025, doi: 10.1371/journal.pone.0324394.

[8] C. Comito, L. Caroprese, and E. Zumpano, "Multimodal fake news detection on social media: a survey of deep learning techniques," Soc. Netw. Anal. Min., vol. 13, no. 1, p. 101, 2023, doi: 10.1007/s13278-023-01104-w.

[9] H. Wang, S. Wang, and Y. H. Han, "Detecting fake news on Chinese social media based on hybrid feature fusion method," Expert Syst. Appl., vol. 208, no. April, p. 118111, 2022, doi: 10.1016/j.eswa.2022.118111.

[10] F. Yan, M. Zhang, and B. Wei, "Multimodal integration for fake news detection on social media platforms," MATEC Web of Conferences, vol. 395, p. 01013, 2024, doi: 10.1051/matecconf/202439501013.

[11] J. Hua, X. Cui, X. Li, K. Tang, and P. Zhu, "Multimodal fake news detection through data augmentation-based contrastive learning," Appl. Soft Comput., vol. 136, Mar. 2023, doi: 10.1016/j.asoc.2023.110125.

[12] K. Nakamura, S. Levy, and W. Y. Wang, "r/Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection," Mar. 2020, [Online]. Available: http://arxiv.org/abs/1911.03854

[13] J. Xie, S. Liu, R. Liu, Y. Zhang, and Y. Zhu, "SeRN: Stance extraction and reasoning network for fake news detection," in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 2520–2524. doi: 10.1109/ICASSP39728.2021.9414787.

[14] S. K. Uppada, P. Patel, and B. Sivaselvan, "An image and text-based multimodal model for detecting fake news in OSN's," J. Intell. Inf. Syst., vol. 61, no. 2, pp. 367–393, 2023, doi: 10.1007/s10844-022-00764-y.

[15] N. C. Hashim, N. A. Abd Majid, and H. Arshad, "Evaluation of multimodal augmented reality experiential learning with emotion detection and speech recognition for learning English vocabulary," E-Proceedings on the Theme of Impact and Influence of Technology & Computing, vol. 70, 2022.

[16] T. Baltrusaitis, C. Ahuja, and L. P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423–443, 2019, doi: 10.1109/TPAMI.2018.2798607.

[17] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal Intelligence: Representation Learning, Information Fusion, and Applications," IEEE Journal on Selected Topics in Signal Processing, vol. 14, no. 3, pp. 478–493, 2020, doi: 10.1109/JSTSP.2020.2987728.

[18] S. Sengan, S. Vairavasundaram, L. Ravi, A. Q. M. Alhamad, H. A. Alkhazaleh, and M. Alharbi, "Fake News Detection Using Stance Extracted Multimodal Fusion-Based Hybrid Neural Network," IEEE Trans. Comput. Soc. Syst., vol. 11, no. 4, pp. 5146–5157, 2024, doi: 10.1109/TCSS.2023.3269087.

[19] R. Kumari, N. Ashok, P. K. Agrawal, T. Ghosal, and A. Ekbal, "Identifying multimodal misinformation leveraging novelty detection and emotion recognition," J. Intell. Inf. Syst., vol. 61, no. 3, pp. 673–694, Dec. 2023, doi: 10.1007/s10844-023-00789-x.

[20] T. Zou, Z. Qian, P. Li, and Q. Zhu, "Pvcg: Prompt-Based Vision-Aware Classification and Generation for Multi-Modal Rumor Detection," ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pp. 11036–11040, 2024, doi: 10.1109/ICASSP48485.2024.10447285.

[21] D. Sharma, S. Purushotham, and C. K. Reddy, "MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain," Sci. Rep., vol. 11, no. 1, Dec. 2021, doi: 10.1038/s41598-021-98390-1.

[22] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings, pp. 457–468, 2016, doi: 10.18653/v1/d16-1044.

[23] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal Factorized Bilinear Pooling with Co-attention Learning for Visual Question Answering," Proceedings of the IEEE International Conference on Computer Vision, vol. 2017-Octob, pp. 1839–1848, 2017, doi: 10.1109/ICCV.2017.202.

[24] N. Xiang, "Deep Learning-Based Fake Information Detection and Influence Evaluation," Comput. Intell. Neurosci., vol. 2022, 2022, doi: 10.1155/2022/8514430.

[25] Y. Guo, H. Ge, and J. Li, "A two-branch multimodal fake news detection model based on multimodal bilinear pooling and attention mtechanism," Front. Comput. Sci., vol. 5, no. April, 2023, doi: 10.3389/fcomp.2023.1159063.

[26] R. Kumari and A. Ekbal, "AMFB: Attention based multimodal Factorized Bilinear Pooling for multimodal Fake News Detection," Expert Syst. Appl., vol. 184, no. March, 2021, doi: 10.1016/j.eswa.2021.115412.

[27] A. Vaswani et al., "Attention Is All You Need," Aug. 2023, [Online]. Available: http://arxiv.org/abs/1706.03762

[28] M. I. Nadeem et al., "EFND: A Semantic, Visual, and Socially Augmented Deep Framework for Extreme Fake News Detection," Sustainability (Switzerland), vol. 15, no. 1, pp. 1–24, 2023, doi: 10.3390/su15010133.

[29] Z. Huang, Y. Hu, Z. Zeng, X. Li, and Y. Sha, "Multimodal Stacked Cross Attention Network for Fine-Grained Fake News Detection," Proc. (IEEE Int. Conf. Multimed. Expo), vol. 2023-July, pp. 2837–2842, 2023, doi: 10.1109/ICME55011.2023.00482.

[30] A. M. Luvembe, W. Li, S. Li, F. Liu, and X. Wu, "CAF-ODNN: Complementary attention fusion with optimized deep neural network for multimodal fake news detection," Inf. Process. Manag., vol. 61, no. 3, p. 103653, 2024, doi: 10.1016/j.ipm.2024.103653.

[31] M. Visweswaran, J. Mohan, S. Sachin Kumar, and K. P. Soman, "Synergistic Detection of Multimodal Fake News Leveraging TextGCN and Vision Transformer," in Procedia Computer Science, Elsevier B.V., 2024, pp. 142–151. doi: 10.1016/j.procs.2024.04.017.

[32] K. Armin, S. Djordje, and Z. Matthias, "Multimodal Detection of Information Disorder from Social Media," in Proceedings - International Workshop on Content-Based Multimedia Indexing, IEEE Computer Society, Jun. 2021. doi: 10.1109/CBMI50038.2021.9461898.

[33] I. Segura-Bedmar and S. Alonso-Bartolome, "Multimodal Fake News Detection," Information (Switzerland), vol. 13, no. 6, 2022, doi: 10.3390/info13060284.

[34] P. Liu, W. Qian, D. Xu, B. Ren, and J. Cao, "Multi-Modal Fake News Detection via Bridging the Gap between Modals," Entropy, vol. 25, no. 4, pp. 1–17, 2023, doi: 10.3390/e25040614.

[35] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, vol. 1, no. Mlm, pp. 4171–4186, 2019.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2016-Decem, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.

[37] L. Ying, H. Yu, J. Wang, Y. Ji, and S. Qian, "Multi-Level Multi-Modal Cross-Attention Network for Fake News Detection," IEEE Access, vol. 9, pp. 132363–132373, 2021, doi: 10.1109/ACCESS.2021.3114093.

[38] I. Musyoka, J. Wandeto, and B. Kituku, "Integrating BERT and RESNET50V2 for Multimodal Cyberbullying Detection∗," 6th International Conference on Intelligent Computing in Data Sciences, ICDS 2024, pp. 1–6, 2024, doi: 10.1109/ICDS62089.2024.10756337.

[39] I. A. Norabid, M. Jalil, R. Ali, A. M. Al-Sbou, and N. H. A. Rahim, "Unifying Modalities: A Comparative Analysis of Bilinear Pooling Fusion Techniques for Multimodal Fake News Detection," Asia-Pacific Journal of Information Technology and Multimedia, vol. 15, no. 1, pp. 349–361, Jun. 2026, doi: 10.17576/apjitm-2026-1501-22.

[40] Y. Yang, L. Zheng, J. Zhang, Q. Cui, Z. Li, and P. S. Yu, "TI-CNN: Convolutional Neural Networks for Fake News Detection," Jan. 2023, [Online]. Available: http://arxiv.org/abs/1806.00749

[41] P. Shaeri and A. Katanforoush, "A Semi-supervised Fake News Detection using Sentiment Encoding and LSTM with Self-Attention," Jul. 2024, [Online]. Available: http://arxiv.org/abs/2407.19332