A Multi-Reading Habits Fusion Adversarial Network for Multi-Modal Fake News Detection

Bofan Wang[1], Shenwu Zhang[2]*
School of Artificial Intelligence, Zhongyuan University of Technology, Zhengzhou, Henan 451191, China[1, 2]
School of Computer Science, Zhongyuan University of Technology, Zhengzhou, Henan 451191, China[1]

***Abstract*—Existing multimodal fake news detection methods face three challenges: the lack of extraction for implicit shared features, shallow integration of multimodal features, and insufficient attention to the inconsistency of features across different modalities. To address these challenges, a multi-reading habits fusion adversarial network for multimodal fake news detection is proposed. In this model, to mitigate the influence of feature changes due to events and emotions, a dual discriminator based on domain adversarial training is built to extract invariant common features. Inspired by the diverse reading habits of individuals, three fundamental reading habits are identified, and a multi-reading habits fusion layer is introduced to learn the interdependencies among the multimodal feature representations of the news. To investigate the semantic inconsistencies of different modalities in news, a similarity constraint reasoning layer is proposed, which first explores the semantic consistency between image descriptions and unimodal features, and then delves into the semantic discrepancies between unimodal and multimodal features. Extensive experimentation has been carried out on the multimodal datasets of Weibo and Twitter. The outcomes indicate that the proposed model surpasses the performance of mainstream advanced benchmarks on both platforms.**

Keywords*—Multimodal fake news detection; feature extraction; feature fusion; consistency alignment

I. INTRODUCTION

In recent years, the rapid growth of social media has significantly reshaped the traditional way people access information. A growing number of users prefer to consume news via social media platforms, these platforms not only ensures the real-time reporting of events from around the world but also provides rich and engaging content in various media forms, such as videos, images, and audio. Compared to simple text reports, news that incorporates images and video elements can convey stories more vividly and thus attract a wider audience. However, this rich medium has also been exploited by fake news, which spreads rapidly through multimedia means. In particular, fake news containing multiple media is more contagious than fake news containing only text, spreading quickly to a wider area and having a more serious impact [1]. Fake news often contains manipulated or completely fabricated images, which are highly misleading and can spread rapidly to a wide audience in a very short time. The spread of fake news can pose a serious threat to public health safety [2] and may affect or even manipulate key political events [3], thus posing a threat to social stability. Therefore, social media platforms urgently need to solve how to quickly and accurately identify fake news.

Based on the content of the news, existing fake news detection technologies are broadly categorized into two groups: unimodal detection methods and multimodal detection methods.

The early focus of unimodal fake news detection was on using feature engineering for artificial feature construction. This includes statistical features such as the frequency of negative vocabulary occurrence and the number of tag symbol repetitions [4], metadata features such as user information, behavioral information, and news platform information [5], language or semantic features of text content [6][7], emotional features of news publishers and content [8], stance features [9][10], writing style and stylistic features [11], content comment features of news [12], and communication based features [13]; The later stage focuses on using static word vector models Word2Vec, Glove, or dynamic word vector models Bert and Roberta to obtain text features.

With the rapid development of social media, the incidence of fake news manipulated through images and text has surged, underscoring the growing importance of detecting multimodal fake news. While strides have been made in multimodal fake news detection technology, several challenges persist:

1) Feature extraction: Current detection methods typically rely on pre-trained models to extract explicit features. For example, using the BERT model to extract text features [14], or using convolutional neural networks such as VGG to extract image features [15]. However, these methods are sensitive to feature distributions. Fake news tends to focus on certain specific fields [15], and these news items usually have negative and pessimistic emotional tones [16][17].

2) Interactive fusion: Existing methods integrate multimodal features to detect fake news through simple early fusion [18] or late fusion [19] strategies, but these fusion strategies are superficial, such as splicing, adding, or simple neural networks to integrate, making it difficult for them to capture the intrinsic dependencies between features.

3) Consistency alignment: Existing methods mainly emphasize capturing similar semantics between different modalities through alignment mechanisms, such as establishing entity alignment [20], relationship alignment [21], and semantic alignment [22] for detection. However, they neglect the acquisition of widely inconsistent semantics.

To tackle these challenges, the Multi-Reading Habits Fusion Adversarial Network (MHFAN) is proposed. MHFAN

*Corresponding Author

aims to detect fake news by extracting both explicit and implicit common features, employing deep feature fusion, and incorporating similarity constraint reasoning.

During the feature extraction stage, the multimodal pre-trained model CLIP is employed to extract explicit features from news text and images. Inspired by the concept of domain adversarial training, adversarial networks are used to construct an event discriminator and an emotion discriminator. This method enables the model to learn features that are insensitive to changes in events and emotions (implicit common features) through adversarial training. Consequently, this approach mitigates the discrepancies in detection results caused by differing event and emotional distributions.

At the feature fusion stage, three common reading habits are identified when people read:

1) Text is the main focus, with images as a supplement: Read the text carefully, but only browse the images briefly.

2) Images are the main focus, with text as a supplement: Observe the images carefully, but quickly skim the text.

3) Equal emphasis on text and images: Read the text carefully and pay the same attention to image details.

To model these three reading habits, the unimodal content initial embeddings of text and images are used to represent the behavior of brief browsing, while the unimodal information is encoded to represent careful reading behavior. Subsequently, a Multi-Reading Habits Fusion Layer (MHF) is designed to simulate the interaction of each reading habit. This layer learns the dependencies between the multimodal features of the news, thereby deepening the feature fusion process.

At the consistency alignment stage, a Similarity Constraint Reasoning Layer (SCR) is proposed to address the inconsistent semantics across different modalities. Initially, a Consistency Reasoning Block (CRB) is constructed to evaluate the consistency between image text descriptions and unimodal features. Subsequently, an Inconsistent Association Constraint (IAC) is applied to quantify the semantic deviation between unimodal and multimodal features.

The main contributions of this work are summarized as follows:

1) In the feature extraction phase, a pre-trained CLIP model was employed to extract explicit features of text and images. Furthermore, a dual discriminator based on the concept of domain adversarial was designed to eliminate the model's dependency on specific events and emotions by extracting implicit features shared across different events and emotional states.

2) Based on people's reading habits, this paper proposes a multimodal feature fusion method that achieves deep integration of different modal features and explores their inconsistencies for the purpose of fake news detection.

3) Explored the Similarity Constraint Reasoning Layer, which can not only measure the consistency between image-expanded semantics and unimodal features but also obtain the semantic deviation between unimodal and multimodal features.

4) Extensive experiments have been conducted on public Weibo and Twitter datasets. The experimental results demonstrate that the MHFAN excels in fake news detection, outperforming conventional detection models across multiple metrics. Additionally, ablation studies have validated the effectiveness of various components within the model, confirming their contributions to the overall performance improvement.

The structure of the remaining sections of this paper is as follows: Section II reviews prior research in the field of content-based fake news detection. Section III provides a detailed introduction to the proposed model and its key components. Section IV describes the datasets utilized, the experimental setup, the baseline models for comparison, and presents the experimental results along with a thorough analysis. Finally, Section V offers a concise summary of the paper.

II. Related Work

Based on the modality of news content, fake news detection methods are categorized into unimodal and multimodal.

Unimodal fake news detection methods primarily encompass three aspects: text-based, vision-based, and metadata-based.

1) Text-based: Early research primarily utilized manual extraction of statistical features from the context of the content. Guo et al. [23]. counted the proportion of negative words in the text, while Parikh et al. [24]. counted the types and numbers of punctuation symbols. However, manual methods are time-consuming and labor-intensive, making it difficult to meet the demands of large volumes of data. As a result, techniques for automatically detecting fake news using deep learning have emerged. Deep neural networks based on CNN [25], RNN [26], attention mechanisms [27], and GNN [6] are constructed to capture semantic, emotional, stylistic, and stance features for identifying fake news.

2) Vision-based: In addition to textual content, some studies also consider image information in the news [27][28]. These methods typically use VGG, ResNet to capture spatial domain features, or through discrete cosine transform, Fourier transform to capture frequency domain features.

3) Metadata-based: The identification of fake news relies not only on the content but also on social contextual features, i.e., metadata. This includes comments [29] (The comment based approach utilizes an interactive mechanism to obtain valuable features between comments and news), user profiles [30] (The method based on user data is suitable for fake news with a large number of users), platform characteristics [27] (Social platform based methods often appear in cross platform fake news detection tasks), and propagation patterns [13] (The method based on propagation mode has time series characteristics). These metadata features are helpful in fake news detection.

Multimodal fake news detection most research focuses on three aspects: feature extraction, interactive fusion, and consistency alignment.

1) Feature extraction: Multimodal news content detection methods typically employ targeted pre-trained models to extract features from different modalities. For instance, word vector models such as Word2Vec and GloVe are used to extract textual features, while convolutional neural networks like VGG and ResNet extract image features [15]. With the advent of transformer architectures [31], Transformer-based pre-trained models have significantly enhanced the capability of feature extraction by capturing deeper linguistic representations.

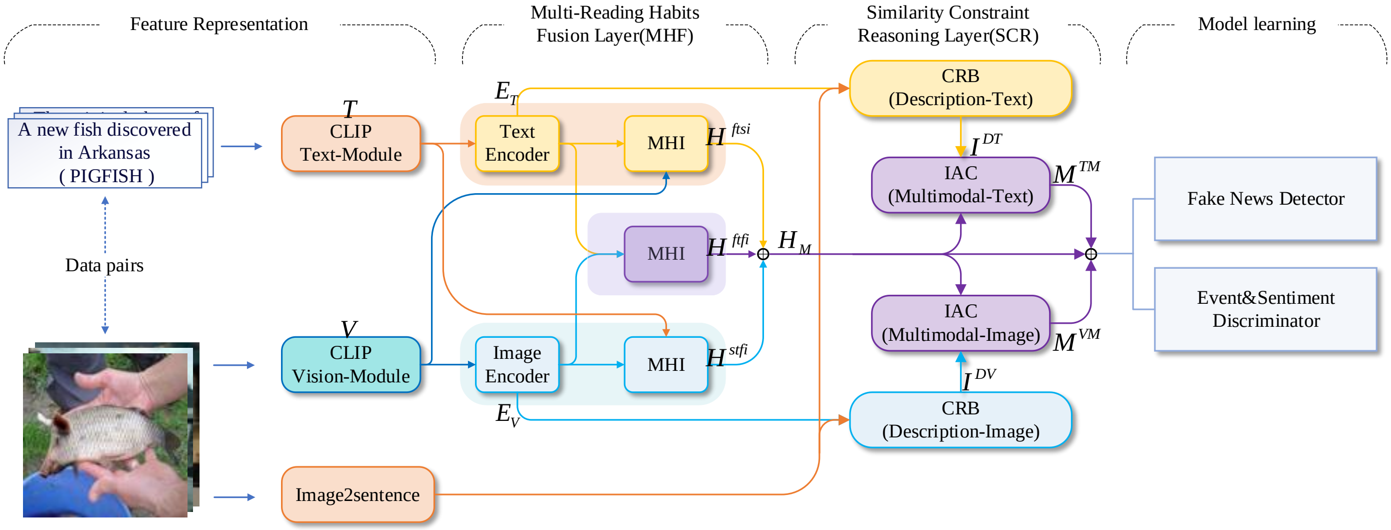


Fig. 1. The model framework MHFAN proposed in this paper consists of four levels: Feature representation, MHF, SCR, and model learning. CLIP and the discriminator are capable of capturing explicit and implicit common features, MHF enhances the deep fusion between features, and SCR can capture the consistency and inconsistency among features.

2) Interactive fusion: The objective of feature interaction fusion is to integrate information from different modality data sources to enhance model performance, and this process is primarily categorized into early fusion and late fusion. Early fusion [32][33][34] also known as feature-level fusion, refers to the combination of different modalities of information through concatenation or addition operations during the feature extraction or feature construction phase of the model. After the fusion, the combined features represent the joint feature space of all modalities, enabling the model to consider information from different data sources simultaneously, with all features being equally output downstream for learning. Late fusion [35][36] also known as decision-level fusion, is where models for each modality are trained independently, learning and extracting features and information from their respective modalities. Each modality's data is first processed and analyzed independently, and the outputs of the same type are fused at the decision stage using operations such as summation, maximum, average, or dot product.

3) Consistency alignment: The mismatch between different information modalities in news is a common source of error, such as discrepancies between images and text. By aligning data from various modalities, we ensure consistency and relevance within a unified representational space. Current research focuses on similarity comparison [20], semantic matching [37], entity alignment [18], and other alignment strategies [38] for detection.

However, the aforementioned methods have the following shortcomings:

- Feature extraction is susceptible to the influence of the distribution of certain news content. For instance, fake news is widely distributed in political and economic spheres and often contains a significant amount of pessimistic emotional content. To mitigate the impact of content bias resulting from data distribution, it is necessary to capture common features that are insensitive to events and emotional changes.
- Methods such as summation, splicing, and averaging for feature fusion are shallow, leading to information loss and redundancy of features;
- Existing alignment methods have not explored the inconsistent information between multimodal features, and there is a lack of correlation and interaction between different types of features.

To address the aforementioned issues, the Multi-Reading Habits Fusion Adversarial Network (MHFAN) has been developed. Initially, during the feature extraction phase, the CLIP pre-trained model is utilized to capture salient features that support the detection task. Concurrently, a dual discriminator is employed to derive common features that are insensitive to event and emotional changes, thereby mitigating bias caused by data distribution. In the feature fusion phase, a Multi-Reading Habits Fusion layer (MHF) is constructed to enhance feature interaction and achieve deep feature integration. Finally, for consistency alignment, a Similarity Constraint Reasoning layer (SCR) is designed to capture both consistencies and inconsistencies between different features, which is then applied to the task of fake news detection.

III. The Proposed Model

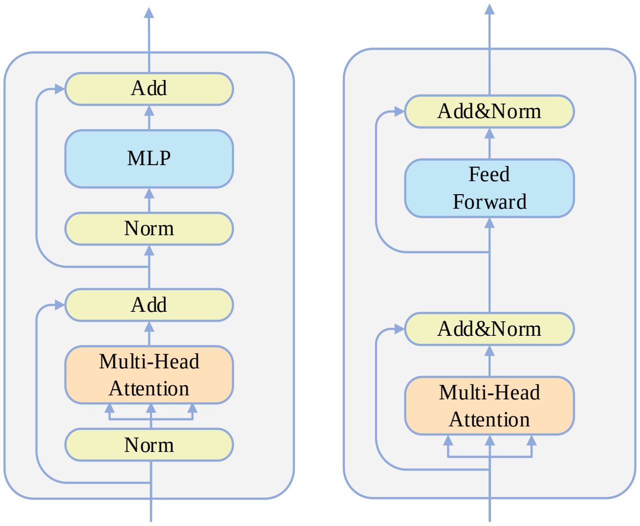
The propose model, MHFAN, has a structure as shown in Fig. 1.

A. *Feature Representation*

The input for MHFAN consists of multimodal news (i.e., image and text content) and image descriptions (expanded from the image content using a pre-trained image2sentence model). For the multimodal news, the text content is represented as a text sequence $T \in \mathbb{R}^{l_T \times d}$, and the sequence T is composed of l_T tokens, where each token $t_i = \mathbb{R}^d$ is a d-dimensional vector learned from the CLIP model. Then, using the CLIP model in the same way to learn the visual features of the image content from the spatial domain, we obtain the image feature vector $V \in \mathbb{R}^{l_V \times d}$ of length l_V in the last hidden layer.

B. *Multi-Reading Habits Fusion Layer(MHF)*

To achieve deep integration of multimodal features, the Multi-Reading Habits Fusion Layer (MHF) has been designed. Based on the differences in how people focus on multimodal information, three reading habits have been identified: "Focus on Image & Scan Text," "Scan Image & Focus on Text," and "Focus on Both Image and Text." In this context, "Focus" indicates thorough reading, while "Scan" indicates cursory reading. Within the MHF, the initial embeddings of the unimodal information are considered as "Scan" behavior, while deeper encoding is regarded as "Focus" behavior. Consequently, MHF first constructs different unimodal encoding blocks and then designs a Multi-Reading Habit Interaction Block (MHI) to model the three types of interactions that occur when people read multimodal information.



(a) Image attention blocks. (b) Text attention blocks.

Fig. 2. The attention block is the basic unit that makes up the encoder.

1) Text and Image encoder: To demonstrate the learning of dependencies between any two text tokens and any two image regions and to extract the intrinsic features of text and images, a Text&Image Encoder based on the self-attention mechanism has been constructed.

The text encoder and the image encoder are attention networks formed by stacking their respective attention blocks, as shown in Fig. 2. The text attention block consists of a multi-head attention mechanism and a feed forward network (FFN), connected through residual connections and layer normalization (Add & Norm). The feed-forward neural network in the image attention block is replaced with a Multilayer Perceptron (MLP). The core of both encoders is the self-attention mechanism, whose computation is illustrated as follows:

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

where, Q, K, and V represent the Query matrix, Key matrix, and Value matrix, respectively. Here, Q=K=V=T, and d_k is equal to $d/2$. To extensively learn richer text contextual information and image upper and lower regional information from different perspectives, the multi-head attention mechanism projects the queries, keys, and values through m different linear projections, and then executes them in parallel. Finally, the processed results are integrated and projected to obtain a new representation, with the computation shown as follows:

$$head = Attention(QW_q, KW_k, VW_v) \qquad (2)$$

$$MultiHead(Q, K, V) = Concat(head_1, \ldots\ldots, head_m) (3)$$

$$E_T \&\& E_V = MultiHead(Q, K, V) \qquad (4)$$

where, $W \in \mathbb{R}^{d \times d}$ are trainable parameters. $E_T \in \mathbb{R}^{l_T \times d}$ is the encoding of the news text content, and $E_V \in \mathbb{R}^{l_V \times d}$ is the encoding of the news image content.

2) Multi-reading habit interaction block (MHI): To model the interactive behaviors within each reading habit, the Multi-Reading Habit Interaction Block (MHI) has been constructed based on the co-attention mechanism to learn the dependencies between multimodal information, as illustrated in Fig. 3. Taking "Focus text&Scan image" as an example, the MHI takes as input the pair $< E_T, V >$, The fusion logic of the MHI is described as follows:

$$\widehat{H_T} = Norm\left(E_T + softmax\left(\frac{E_T(V)^T}{\sqrt{d}}\right)V\right) \qquad (5)$$

$$\widehat{H_V} = Norm\left(V + softmax\left(\frac{V(E_T)^T}{\sqrt{d}}\right)E_T\right) \qquad (6)$$

$$H_T^{ftsi} = Norm(\widehat{H_T} + FFN(\widehat{H_V})) \qquad (7)$$

$$H_V^{ftsi} = Norm(\widehat{H_V} + FFN(\widehat{H_T})) \qquad (8)$$

$$H^{ftsi} = concat(H_T^{ftsi}, H_V^{ftsi}) \qquad (9)$$

H^{ftsi}represents the integrated semantics of the interaction block specifically for the "Focus text&Scan image" reading habit. The integrated semantics for the "Focus image&Scan text" and "Focus image&Focus text" reading habits are H^{fist} and H^{fift}, respectively.

Finally, the three reading habits are integrated to form a comprehensive fused representation of the multimodal news, denoted as $H_M = concat(H^{ftsi}, H^{ftfi}, H^{fist})$.

C. Similarity Constraint Reasoning Layer (SCR)

To explore the consistency and inconsistency between different modal features, the **S**imilarity **C**onstraint **R**easoning (**SCR**) layer has been designed from two perspectives. First, the **C**onsistency **R**easoning **B**lock (**CRB**) is employed to investigate the consistency between image descriptions and unimodal features. Then, the **I**nconsistent **A**ssociation **C**onstraint (**IAC**) is introduced to capture the semantic deviations between unimodal features and multimodal fused features.

1) Consistency Reasoning Block (CRB): Taking the consistency alignment between the image description S and the textual features E_T as an example, S and E_T are first projected into a shared latent space of the same dimension.

$$F_S = \tanh(W_c S + b_c) \tag{10}$$

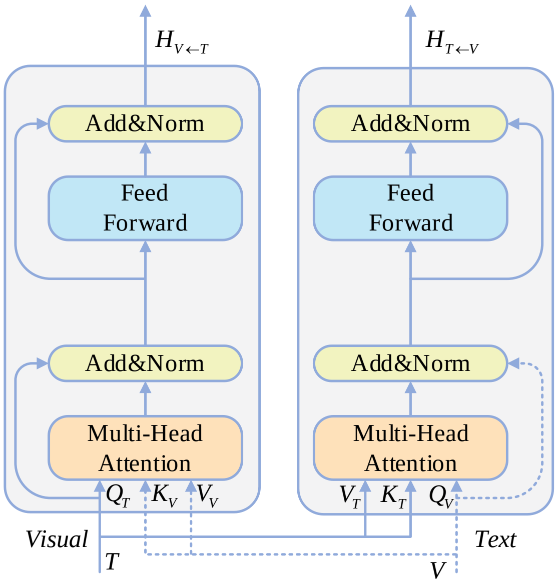
$$F_T = \tanh(W_m E_T + b_m) \tag{11}$$

where, $F_S$and F_T represent the image description and the deep textual semantics in the shared space, respectively. Then, the image description $Q_s = WqF_S$ is used as the query and the deep textual feature $K_T = W_k F_T$ as the keys. Through the attention weight matrix $A_{CM} = softmax(Q_c K_T^T)$, the consistency between these two features is captured. The matrix A_{CM} reflects the degree of attention that the query vector Q_s pays to the key vector K_T.

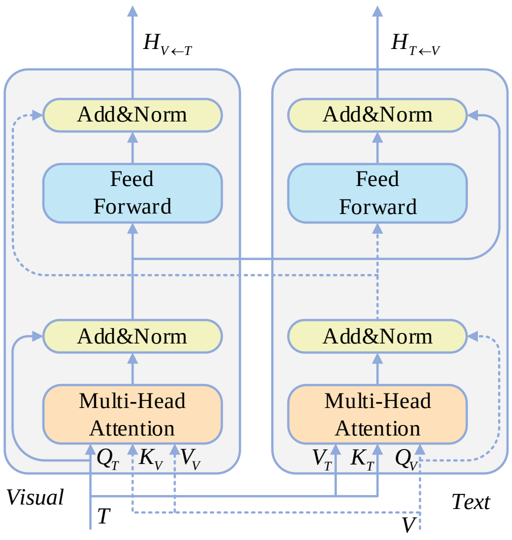
$$A_{CM} = softmax(Q_c K_T^T) \tag{12}$$

$$I^{ST} = F_s + A_{CM} F_T \tag{13}$$

where, I^{ST}represents the consistency aggregation vector between the description and the text, and the consistency aggregation vector between the description and the vision, I^{SV}, follows the aforementioned equation.



(a) Traditional co-attention mechanism.



(b) Multi-Reading Habit Interaction Block (MHI).

Fig. 3. The architecture diagram of co-attention and our MHI.

2) Inconsistent Association Constraint(IAC): The **I**nconsistent **A**ssociation **C**onstraint (**IAC**) is designed to measure the semantic deviation between unimodal and multimodal information in the news. It assesses the deviation between the unimodal aggregated vectors (I^{ST} and I^{SV}) and the multimodal fused semantics. Taking the deviation between I^{ST} and H^M as an example:

$$M_{i,j}^{TM} = \cos(I_i^{ST}, H_j^M) \tag{14}$$

where $M_{i,j}^{TM} \in \mathbb{R}^{l_{ST} \times l_M}$, l_{ST} and l_M are the lengths of the list I^{ST} and H^Mrespectively, and$M^{TM}$is the text multimodal deviation matrix. In this way, M^{VM} represents the image multimodal deviation between I^{SV} and H^M. Subsequently, the two types of deviation matrices are passed to an MLP to obtain the overall semantic deviation M^{all}:

$$M^{all} = MLP(concat(M^{TM}, M^{VM})) \tag{15}$$

Thus, the final multimodal fused features of the news are the comprehensive measure IM of the multimodal features, the consistency aggregated vectors, and the overall semantic deviation:

$$IM = concat(H^M, I^{SV}, I^{ST}, M^{all}) \tag{16}$$

D. Model Learning

The model learning is accomplished by the Fake News Detector and the Event&Sentiment Discriminator. The former consists of a fully connected layer and a softmax layer, with the purpose of correctly classifying the news; the latter is composed of a gradient reversal layer (GRL) and a fully connected layer, with the aim of accurately classifying the news events and sentiments. Both use cross-entropy to calculate the loss. The loss function L_f for the Fake News Detector is defined as follows:

$$L_f = -[y_f \log P_f + (1 - y_f)\log(1 - P_f)] \tag{17}$$

where, P_f represents the predicted label, and y_f is the true label. Similarly, the loss functions for the Event Discriminator and the Sentiment Discriminator are L_e and L_s, respectively.

In fact, due to the presence of the GRL, the Discriminator is inclined to maximize the loss function. A higher loss indicates that the feature distributions are similar, which eliminates the dependency on specific events or specific sentiments. The features learned are common across different events or different sentiments. This sets up a minimax game with the Detector, which tends to minimize the objective function, establishing an adversarial relationship. The final loss function for the model is defined as:

$$L_{final} = L_f - \alpha L_e - \beta L_s \tag{18}$$

where, the loss function parameters α and β are used to balance the losses between fake news detection and event and sentiment classification.

IV. Experiments

A. Datasets and Data Preprocessing

To verify the performance of MHFAN, experiments were conducted on two datasets: Weibo and Twitter. The Weibo dataset, proposed by Jin et al [39]., includes confirmed fake news verified by Sina Weibo's official platform from May 2012 to January 2016, as well as real news verified by Xinhua, an authoritative Chinese news source. The Twitter dataset, proposed by Boididou et al. [40], is used to evaluate multimodal tasks on MediaEval. During the data preprocessing phase, duplicate images were removed, low-quality images were filtered out, and punctuation, numbers, special characters, and short words were eliminated from the text.

It was observed that in both the Weibo and Twitter multimodal datasets, the images and their corresponding text content were not entirely relevant and lacked some semantic information to varying degrees. To address this issue, a pre- trained image2sentence model [41] was employed to generate brief descriptions of the images, thereby providing text information that aligns with the image content. This generated text was used to expand the textual content of the dataset and fill in the missing semantic information. Additionally, the SKEP model [42] was utilized to categorize the datasets into three emotional labels (positive, neutral, negative), and the Single-Pass method [43] was used to detect new events mentioned in the posts.

B. Experimental Settings

To prevent overfitting, the model parameters of CLIP were frozen during training on both the Twitter and Weibo datasets. In the embedding layer, the length of the text sequence was set to 128, and the length of the image representation was 197; the Text&Image Encoder had six attention heads and consisted of 4 attention blocks; furthermore, the model was trained for 100 epochs with a learning rate of 1e-5, and the batch size was set to 128.

C. Evaluation Metrics

The experiments utilized accuracy, precision, recall, and the F1 score to assess the performance of the proposed model.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{19}$$

$$Precision = \frac{TP}{TP+FP} \tag{20}$$

$$Recall = \frac{TP}{TP+FN} \tag{21}$$

TABLE I. COMPARISON RESULTS OF MHFAN WITH DIFFERENT BASELINE MODELS ON THESE TWO DATASETS

Dataset	**Methods**	**Accuracy**	**Fake News**			**True News**		
			Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Twitter	Visual-Only	0.590	0.580	0.540	0.560	0.600	0.640	0.620
	Text-Only	0.529	0.488	0.497	0.496	0.565	0.556	0.561
	Att-RNN	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN	0.648	0.810	0.498	0.617	0.584	0.759	0.660
	MVAE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	MCAN	0.809	0.889	0.765	0.822	0.732	0.871	0.795
	MEAN	0.780	0.690	**0.840**	0.760	**0.870**	0.740	0.800
	MHFAN	**0.840**	**0.924**	0.813	**0.865**	0.736	**0.887**	**0.804**
Weibo	Visual-Only	0.640	0.580	0.570	0.610	0.640	0.690	0.660
	Text-Only	0.640	0.741	0.573	0.646	0.651	0.798	0.711
	Att-RNN	0.772	0.854	0.656	0.742	0.720	0.889	0.795
	EANN	0.782	0.827	0.697	0.756	0.752	0.863	0.804
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	MCAN	0.899	**0.913**	0.889	0.901	0.884	**0.909**	0.897
	MEAN	0.894	0.900	0.870	0.890	0.890	0.910	0.900
	MHFAN	**0.905**	0.887	**0.931**	**0.909**	**0.925**	0.877	**0.901**

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{22}$$

where, TP: fake news forecast is fake; TN: real news predicted to be real; FP: real news that is predicted to be fake; FN: fake news predicted to be real.

D. Performance Comparison

MHFAN was compared with several advanced baselines, including both unimodal and multimodal models.

1) Unimodal Models:

a) Visual-Only: This model relies entirely on image information for subsequent classification, using a pre-trained VGG-19 model to extract image features.

b) Text-Only: This model relies entirely on text information for subsequent classification, using Word2Vec combined with Text-CNN to extract text features.

Multimodal Models:

c) Att-RNN [39]: It employs a cross-modality attention mechanism to combine text, visual, and social context features.

d) EANN [15]: While using pre-trained models to extract explicit features from the display, it constructs an event recognizer to obtain implicit common features.

e) MVAE [44]: It uses an encoding-decoding paradigm to capture shared representations that include both visual and textual modalities.

f) MCAN [14]: It integrates features from text, spatial domain, and frequency domain through deeply stacked co-attention layers.

TABLE II. ABLATION ANALYSIS ON TWITTER AND WEIBO DATASETS

Dataset	Methods	Accuracy	Fake News			True News		
			Precision	Recall	F1	Precision	Recall	F1
Twitter	-Text	0.593	0.744	0.539	0.625	0.467	0.685	0.555
	-Image	0.678	0.789	0.666	0.723	0.552	0.699	0.617
	-Description	0.762	0.854	0.750	0.798	0.648	0.783	0.710
	-MHF	0.774	0.842	0.788	0.814	0.676	0.748	0.710
	-SCR	0.807	0.859	0.830	0.844	0.727	0.768	0.747
	-Adversarial	0.819	0.905	0.796	0.847	0.712	0.858	0.778
	MHFAN	**0.840**	**0.924**	**0.813**	**0.865**	**0.736**	**0.887**	**0.804**
Weibo	-Text	0.635	0.640	0.642	0.642	0.630	0.628	0.629
	-Image	0.669	0.673	0.681	0.677	0.666	0.657	0.661
	-Description	0.775	0.782	0.773	0.777	0.768	0.778	0.772
	-MHF	0.773	0.794	0.750	0.771	0.755	0.798	0.776
	-SCR	0.805	0.817	0.793	0.805	0.793	0.816	0.805
	-Adversarial	0.858	0.859	0.865	0.861	0.858	0.852	0.856
	MHFAN	**0.905**	**0.887**	**0.931**	**0.909**	**0.925**	**0.877**	**0.901**

g) MEAN [45]: It utilizes a multimodal generator to enhance the latent discriminative feature representations of text and image modalities.

The results are shown in Table I from which the following observations can be made:

- Compared to the Visual-Only approach that solely relies on visual modality, Text-Only demonstrates a distinct advantage in the task of fake news detection. This suggests that visual information has relatively limited expressiveness and struggles to provide semantic information as rich as text. Therefore, in the task of fake news detection, the textual modality has been proven to be more effective than the visual modality, and better capable of distinguishing between true and fake news information.
- Att-RNN achieves better performance than Visual-Only and Text-Only, indicating that the application of multimodal information is beneficial for detection; EANN constructs an event adversarial neural network and demonstrates strong performance in fake news detection tasks using explicit and implicit common features; MVAE surpasses EANN and Att-RNN in fake news detection with the superior performance of its multimodal variational autoencoder; MEAN improves the model's performance by capturing and learning common features of modalities and events through dual discriminators; MCAN's designed co-attention network shows superior performance compared to MVAE and MEAN, indicating the effectiveness of capturing consistent semantics between multimodal features.
- Compared with the comparative model, the proposed MHFAN fake news detection model shows superiority in various indicators on Weibo and Twitter datasets. In the Twitter dataset, the accuracy of fake news detection increased by 3.5%. On the Weibo dataset, the recall rate of fake news detection increased by 4.2%.

E. Ablation Analysis

1) Effectiveness of each component: To investigate the effectiveness of each component in MHFAN, five model variants were created: -Text, -Image, -Description, -MHF, -SCR, and -Adversarial. These variants denote the removal of the following components: text representation, image representation, image description, the MHF, the SCR, and the adversarial network, respectively.

The results of the ablation study are shown in TABLE II. , from which the following observations can be made:

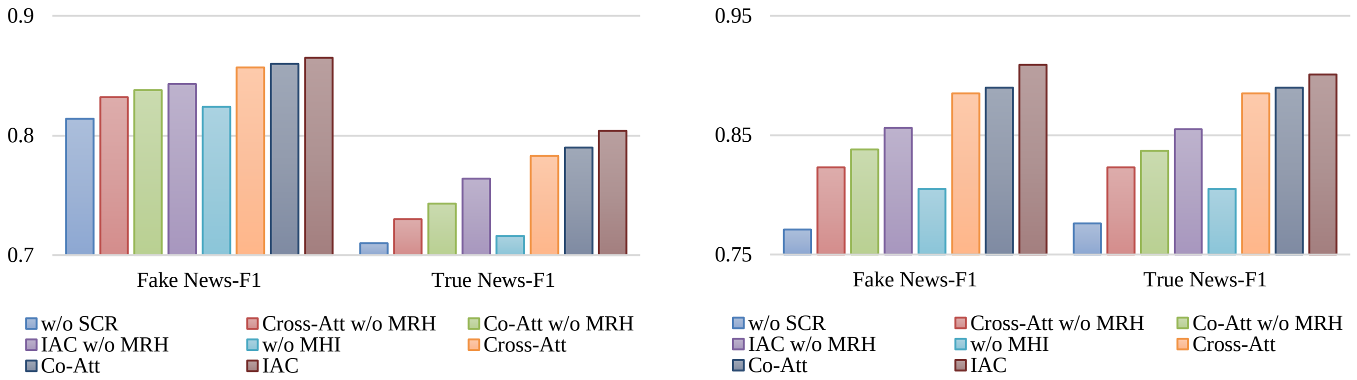
- The removal of different layers led to varying degrees of degradation, demonstrating the effectiveness of each component.
- The -Text and -Image models performed weaker than MHFAN, confirming that relying solely on unimodal information is detrimental to detection. The -Description model also saw a significant performance drop, indicating the importance of image descriptions for semantic expansion.
- The performance of MHFAN without the MHF layer was significantly reduced, reflecting that modeling human reading habits can promote the tight integration of multimodal information; the absence of the SCR layer meant that MHFAN could not obtain the consistency and inconsistency of features between modalities, and its performance also plummeted.
- The -Adversarial model experienced a decrease in precision of 1.9% and 2.8% on the Twitter and Weibo datasets, respectively, illustrating the importance of capturing implicit common features for fake news detection.

2) Comparative analysis of multi-reading habits fusion layer: MHF is the method employed by MHFAN for deep feature fusion and includes two core mechanisms: Multi-Reading Habits (MRH) and the Multi-Reading Habit Interaction Block (MHI). The MRH captures both deep and shallow features of different modalities, while the MHI achieves interactive fusion of features from these modalities. Comparative experiments were conducted under two conditions: one with MRH and one without MRH (w/o MRH). Three alternative methods were also tested to replace MHI: traditional co-attention [46], cross-attention [47], and a version without the MHI module (w/o MHI).

In 0, it is observed that the removal of either MRH or MHI significantly degrades the performance of MHFAN. Under both conditions with MRH and without MRH (w/ MRH and w/o MRH), cross-att, co-att, and IAC all demonstrate superior performance compared to those without SCR (w/o SCR), indicating the necessity for deep interaction between features of different modalities. Moreover, SCR outperforms the other three alternative methods, suggesting that SCR can enhance the interaction of each reading habit, thereby achieving a deeper multimodal feature fusion. Additionally, whether it is the alternative methods or IAC, the scenarios with MRH (w/ MRH) show better results than without MRH (w/o MRH), demonstrating the importance of capturing deep and shallow features from different modalities.

3) Comparative analysis of inconsistent association constraints: Several alternative methods to IAC within MHFAN were evaluated by replacing IAC, which captures feature inconsistency, with the following methods: -IAC (removal of the IAC module), KL-divergence, Euclidean distance, Orthogonality constraints [48], and RA-coherence [49].

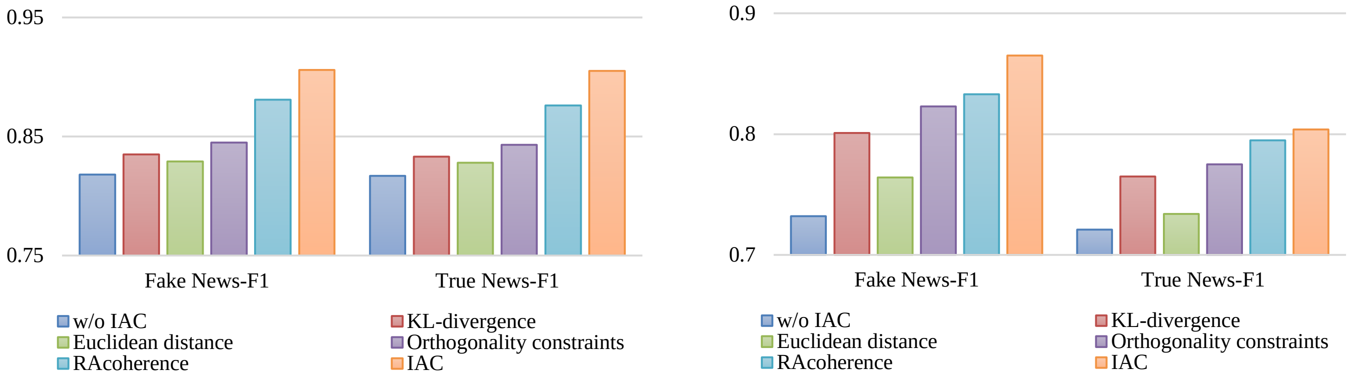
The results, as shown in 0, indicate that compared to w/o IAC, all four variants perform better on both datasets, demonstrating the importance of capturing semantic deviations between different modalities in multimodal fake news detection. Furthermore, IAC outperforms the four alternative methods on both datasets as well. IAC captures the inconsistency between news modalities by calculating the correlation matrices of the two modalities, while the other four methods focus more on the correlation between two types of features and lack an effective measurement of different feature distributions. This proves the superiority of IAC in handling semantic deviations.



(a) On the Twitter dataset

(b) On the Weibo dataset

Fig. 4. Comparison of performance of different ablation blocks in MHF.



(a)On the Twitter dataset

(b) On the Weibo dataset

Fig. 5. Comparison of performance of different ablation blocks in IAC.

V. Case Study

To further illustrate the effectiveness of the proposed method, several cases from the Twitter dataset were selected for visualization.

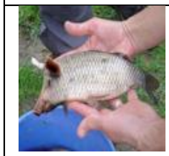
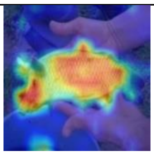
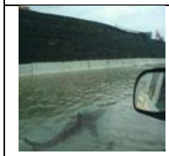
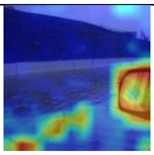
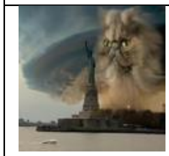
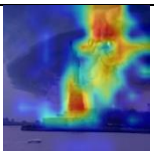
Image	Visualization	Text	Image description
		A new fish discovered in Arkansas (PIGFISH)	A fish with a head resembling a pig's face.
		Casual shark swimming passed a car	There is a shark swimming in the water next to the car mirror.
		The original photo of Hurricane Sandy over the Statue of Liberty	There is a storm behind the Statue of Liberty that looks like a cat face.

Fig. 6. Visualization case of twitter dataset.

From Fig. 6, it is evident that MHFAN effectively captures features within images and aligns semantically with the corresponding text and image descriptions. In the first example, MHFAN identified the body of the fish and the face of a pig in the image, even though the text did not directly mention pigs, underscoring the importance of image descriptions for semantic expansion. In the second example, MHFAN adeptly focused on the shark and the rearview mirror in the image, achieving semantic alignment with both the text and the image description. The third example similarly demonstrates the model's strengths in feature extraction and consistent semantic alignment.

VI. Conclusion

The dissemination of fake news not only undermines the credibility of news media but also negatively affects the online information environment. The spread of false information severely impedes the healthy development of social media platforms. In response to the existing issues in multimodal fake news detection, the Multi-Reading Habits Fusion Adversarial Network (MHFAN) has been developed, and its effectiveness has been extensively tested and verified on two datasets. Future work aims to refine MHFAN by incorporating factual data from search engines and metadata associated with news articles. This enhancement strategy is expected to bolster the network's resilience and expand its applicability to a wider range of scenarios.

Acknowledgment

This research was supported by the Key Scientific Research Project of Colleges and Universities in Henan Province, titled "Research on Key Technologies for Text Processing in Network Public Opinion Monitoring" (Project No. 22B520054).

References

[1] D. S. Nielsen, R. McConville, “MuMiN: A large-scale multilingual multimodal fact-checked misinformation social network dataset,” Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval, pp. 3141-3153, 2022.

[2] K. Roitero, M. Soprano, B. Portelli, D. Spina, V. Della Mea, G. Serra, “The COVID-19 infodemic: Can the crowd judge recent misinformation objectively?” Proceedings of the 29th ACM International Conference on Information and Knowledge Management, pp. 1305-1314, 2020.

[3] M. Osmundsen, A. Bor, P. B. Vahlstrup, A. Bechmann, M. B. Petersen, “Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter,” American Political Science Review, vol. 115, no. 3, pp. 999-1015, 2021.

[4] C. Castillo, M. Mendoza, B. Poblete, “Information credibility on Twitter,” In Proceedings of the 20th International Conference on World Wide Web, pp. 675-684, 2011.

[5] S. Wu, Q. Liu, Y. Liu, L. Wang, T. Tan, “Information Credibility Evaluation on social media,” In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, pp. 4403-4404, 2016.

[6] L. Hu, T. Yang, L. Zhang, W. Zhong, D. Tang, C. Shi, "Compare to The Knowledge: Graph Neural Fake News Detection with External Knowledge," In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp. 754-763, 2021.

[7] L. Wu, R. Yuan, L. Sun, W. He, "Evidence Inference Networks for Interpretable Claim Verification," Proceedings of the AAAI Conference on Artificial Intelligence, pp. 14058-14066, 2021.

[8] X. Zhang, J. Cao, X. Li, Q. Sheng, L. Zhong, K. Shu, "Mining Dual Emotion for Fake News Detection," In Proceedings of the Web Conference 2021, pp. 3465-3476, 2021.

[9] L. Wu, Y. Rao, H. Jin, A. Nazir, L. Sun, "Different Absorption from the Same Sharing: Sifted Multi-task Learning for Fake News Detection," In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 4644-4653, 2019.

[10] J. Xie, S. Liu, R. Liu, Y. Zhang, Y. Zhu, "SERN: Stance Extraction and Reasoning Network for Fake News Detection," In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2520-2524, 2021.

[11] L. Wu, Y. Rao, C. Zhang, Y. Zhao, A. Nazir, "Category-Controlled Encoder-Decoder for Fake News Detection," IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 2, pp. 1242-1257, 2023.

[12] K. Shu, L. Cui, S. Wang, D. Lee, H. Liu, "dEFEND: Explainable Fake News Detection," In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 395-405, 2019.

[13] K. Shu, D. Mahudeswaran, S. Wang, "Hierarchical Propagation Networks for Fake News Detection: Investigation and Exploitation," In Proceedings of the International AAAI Conference on Web and Social Media, pp. 626-637, 2020.

[14] Y. Wu, P. Zhan, Y. Zhang, L. Wang, Z. Xu, "Multimodal Fusion with Co-Attention Networks for Fake News Detection," In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 2560-2569, 2021.44

[15] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, "EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection," In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 849-857, 2018.99

[16] O. Ajao, D. Bhowmik, S. Zargari, "Sentiment Aware Fake News Detection on Online Social Networks," In ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2507-2511, 2019.66

[17] A. Giachanou, P. Rosso, F. Crestani, "Leveraging Emotional Signals for Credibility Detection," In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 877-880, 2019.77

[18] P. Li, X. Sun, H. Yu, Y. Tian, F. Yao, G. Xu, "Entity-Oriented Multi-Modal Alignment and Fusion Network for Fake News Detection," IEEE Transactions on Multimedia, vol. 24, pp. 3455-3468, 2022.11

[19] X. Zhou, J. Wu, R. Zafarani, "Similarity-Aware Multi-modal Fake News Detection," In Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, pp. 354-367, 2020.22

[20] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, "Cross-modal Ambiguity Learning for Multimodal Fake News Detection," In Proceedings of the ACM Web Conference 2022, pp. 2897-2905, 2022.33

[21] M. Dhawan, S. Sharma, A. Kadam, R. Sharma, P. Kumaraguru, "Game-on: Graph Attention Network Based Multimodal Fusion for Fake News Detection," Social Network Analysis and Mining, vol. 14, pp. 1-13, 2022.55

[22] Y. Ganin, V. Lempitsky, "Unsupervised Domain Adaptation by Backpropagation," In Proceedings of the 32nd International Conference on Machine Learning, pp. 1180-1189, 2015.88

[23] C. Guo, J. Cao, X. Zhang, K. Shu, M. Yu, "Exploiting Emotions for Fake News Detection on Social Media," ArXiv, vol. abs/1903.01728, 2019.

[24] S. B. Parikh, P. K. Atrey, "Media-Rich Fake News Detection: A Survey," In Proceedings of the 2018 IEEE Conference on Multimedia Information Processing and Retrieval, pp. 436-441, 2018.

[25] P. K. Verma, P. Agrawal, I. Amorim, R. Prodan, "WELFake: Word Embedding Over Linguistic Features for Fake News Detection," IEEE Transactions on Computational Social Systems, vol. 8, no. 4, pp. 881-893, 2021.

[26] L. Wu, Y. Rao, Y. Zhao, H. Liang, A. Nazir, "DTCA: Decision Tree-based Co-Attention Networks for Explainable Claim Verification," In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1024-1035, 2020.

[27] P. Qi, J. Cao, T. Yang, J. Guo, J. Li, "Exploiting Multi-domain Visual Information for Fake News Detection," In Proceedings of the 2019 IEEE International Conference on Data Mining, pp. 518-527, 2019.

[28] S. Abdali, R. Gurav, S. Menon, D. Fonseca, N. Entezari, N. Shah, "Identifying Misinformation from Website Screenshots," Proceedings of the International AAAI Conference on Web and Social Media, vol. 15, no. 2, pp. 13-23, 2021.

[29] H. Choi, Y. Ko, "Using Topic Modeling and Adversarial Neural Networks for Fake News Video Detection," In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 2950-2954, 2021.

[30] Y. Dou, K. Shu, C. Xia, P. S. Yu, L. Sun, "User Preference-aware Fake News Detection," In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2051-2055, 2021.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, “Attention is all you need,” Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000-6010, 2017.

[32] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, S. Shin'ichi, "SpotFake: A Multi-modal Framework for Fake News Detection," In Proceedings of the 2019 IEEE Fifth International Conference on Multimedia Big Data, pp. 39-47, 2019.

[33] S. Y. Boulahia, A. Amamra, M. R. Madi, S. Daikh, "Early, Intermediate and Late Fusion Strategies for Robust Deep Learning-based Multimodal Action Recognition," Mach. Vision Appl, vol. 32, no. 6, pp. 1-18, Nov 2021.

[34] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, L. Wei, "Detecting Fake News by Exploring the Consistency of Multimodal Data," Inf. Process. Manage, vol. 58, no. 5, pp. 1-13, 2021.

[35] P. Meel, D. K. Vishwakarma, "Multi-modal Fusion Using Fine-tuned Self-attention and Transfer Learning for Veracity Analysis of Web Information," Expert Syst. Appl., pp. 1-16, 2023.

[36] S. Singhal, M. Dhawan, R. R. Shah, P. Kumaraguru, "Inter-modality Discordance for Multimodal Fake News Detection," In Proceedings of the 3rd ACM International Conference on Multimedia in Asia, pp. 33, 2022.

[37] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, L. Wei, "Detecting Fake News by Exploring the Consistency of Multimodal Data," Inf. Process. Manage., vol. 58, no. 5, pp. 1-13, 2021.

[38] P. Qi, J. Cao, X. Li, H. Liu, Q. Sheng, X. Mi, "Improving Fake News Detection by Using an Entity-enhanced Framework to Fuse Diverse Multimodal Clues," In Proceedings of the 29th ACM International Conference on Multimedia, pp. 1212-1220, 2021.

[39] Z. Jin, J. Cao, H. Guo, Y. Zhang, J. Luo, "Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs," In Proceedings of the 25th ACM International Conference on Multimedia, pp. 795-816, 2017.

[40] C. Boididou, S. Papadopoulos, D. T. Dang-Nguyen, G. Boato, M. Riegler, S. Middleton, "Verifying Multimedia Use at MediaEval 2016," In MediaEval Benchmarking Initiative for Multimedia Evaluation, 2015.

[41] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, "Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 652-663, 2017.

[42] H. Tian, C. Gao, X. Xiao, H. Liu, B. He, H. Wu, "SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis," In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4067-4076, 2020.

[43] Z. Jin, J. Cao, Y.-G. Jiang, Y. Zhang, "News Credibility Evaluation on Microblog with a Hierarchical Propagation Model," In Proceedings of the 2014 IEEE International Conference on Data Mining, pp. 230-239, 2014.

[44] D. Khattar, J. S. Goud, M. Gupta, V. Varma, "MVAE: Multimodal Variational Autoencoder for Fake News Detection," In Proceedings of the World Wide Web Conference, pp. 2915-2921, 2019.

[45] P. Wei, F. Wu, Y. Sun, H. Zhou, X.-Y. Jing, "Modality and Event Adversarial Networks for Multi-Modal Fake News Detection," IEEE Signal Processing Letters, pp. 1382-1386, 2022.

[46] Lu, J., Batra, D., Parikh, D., Lee, S. "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks," In Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp. 2-12, 2019.

[47] Chen, C.-F. R., Fan, Q., Panda, R. "CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification," In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, pp. 347-356, 2021.

[48] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, D. Erhan, "Domain Separation Networks," In Proceedings of the 30th International Conference on Neural Information Processing Systems, pp. 343-351, 2016.

[49] W. Zhang, W. Lam, Y. Deng, J. Ma, "Review-guided Helpful Answer Identification in E-commerce," In Proceedings of The Web Conference, pp. 2620-2626, 2020.