Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.047053

ARTICLE

Fake News Detection Based on Text-Modal Dominance and Fusing Multiple Multi-Model Clues

Lifang Fu[1], Huanxin Peng[2,*], Changjin Ma[2] and Yuhan Liu[2]

[1]College of Arts and Sciences, Northeast Agricultural University, Harbin, 150030, China

[2]College of Engineering, Northeast Agricultural University, Harbin, 150030, China

*Corresponding Author: Huanxin Peng. Email: s210701005@neau.edu.cn

Received: 23 October 2023 Accepted: 24 January 2024 Published: 26 March 2024

ABSTRACT

In recent years, how to efficiently and accurately identify multi-model fake news has become more challenging. First, multi-model data provides more evidence but not all are equally important. Secondly, social structure information has proven to be effective in fake news detection and how to combine it while reducing the noise information is critical. Unfortunately, existing approaches fail to handle these problems. This paper proposes a multi-model fake news detection framework based on Tex-modal Dominance and fusing Multiple Multi-model Cues (TD-MMC), which utilizes three valuable multi-model clues: text-model importance, text-image complementary, and text-image inconsistency. TD-MMC is dominated by textural content and assisted by image information while using social network information to enhance text representation. To reduce the irrelevant social structure's information interference, we use a unidirectional cross-modal attention mechanism to selectively learn the social structure's features. A cross-modal attention mechanism is adopted to obtain text-image cross-modal features while retaining textual features to reduce the loss of important information. In addition, TD-MMC employs a new multi-model loss to improve the model's generalization ability. Extensive experiments have been conducted on two public real-world English and Chinese datasets, and the results show that our proposed model outperforms the state-of-the-art methods on classification evaluation metrics.

KEYWORDS

Fake news detection; cross-modal attention mechanism; multi-modal fusion; social network; transfer learning

1 Introduction

In today's era, information spreads more quickly and promptly on social media websites with the Internet's rapid development. Everyone can create information and influence public opinion, and fake news arises [1]. Fake news spreaders often deliberately manipulate, falsify, or exaggerate original information to mislead users, causing adverse impacts on the public and society in the absence of proper monitoring and suppression. For example, there was a time when it was widely spread on the Weibo platform that Shuanghuanglian oral liquid had an *in vitro* inhibitory effect on the novel coronavirus. Many people frantically bought related drugs following the crowd, causing serious market

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

chaos. Some residents even got infected or passed the virus to others when they went out to buy medicine [2]. So effectively detecting fake news has become a required task.

Existing studies have achieved remarkable results in extracting textual features [3]. For example, Ma et al. captured content semantics and propagation structure by a recurrent neural network (RNN) for rumor detection [4]. Meel et al. proposed a semi-supervised temporal ensembling-based convolutional neural network to identify fake news. In this model, linguistic and stylistic information features of annotated news articles are extracted by ConvNet filters [5]. In recent years, information has diversified, and people are increasingly inclined to combine visual and text content to express their ideas and emotions [6]. These multi-modal data can provide more evidence but also add many challenges. Early works in multi-model fake news detection mainly extracted features from each modality and implemented multi-model fusion by concatenation operations. Singhal et al. [7] extracted textual and visual information using Bidirectional Encoder Representation from Transformers (BERT) and Visual Geometry Group (VGG-19) and then combined them to detect fake news. Wu et al. [8] developed a multi-model co-attention network to learn the interdependence between text and images. However, the simple fusion does not capture the high level of cross-model features and even important information is ignored, affecting model performance [9]. Some work attempts to construct the correlation between models (such as semantic alignment and entity alignment) to fuse multi-model features. For example, Hu et al. [10] constructed a mutual learning network to learn the potential consistency of text and vision. Qi et al. [11] fully used three valuable text-image correlation properties: entity inconsistency, entity mutual enhancement, and text complementary relationship to detect fake news. Qian et al. [12] studied inter-modal and intra-modal relationships of text and images through a multi-model contextual attention network (HMCAN). However, we discover that multi-model fake news still faces three major problems. (1) Modeling the multi-model content insufficiently: there are inconsistencies between the text and images and exploring the mismatch characteristics can easily lead to the conclusion that the news is fake. Additionally, social network information can facilitate fake news detection which is often ignored in existing methods [13]. (2) Fusion noise: multi-model feature fusion can accumulate irrelevant information, adding noise to the original content. (3) While multi-modal information is conducive to improving the performance of fake news detection, not all models are equally important.

To solve the above problems, a novel multi-model fake news detection framework based on Text-modal Dominance and fusing Multiple Multi-model Cues (TD-MMC) is proposed, which can effectively combine textual, visual, and social graph features in one unified framework to learn three multi-model clues and reduce fusion noise. Specifically, TD-MMC first extracts semantic-level vector representations of the textual, visual, and social networks and uses multi-head self-attention to enhance three intra-modal features. Next, a two-round cross-modal fusion is adopted, which uses social networks to supplement the original text to obtain enhanced textural features and then learns enhanced textural text and image cross-model fusion feature representation. It also captures the inconsistent information of enhanced text and image features. Finally, enhanced textual features, cross-model fusion features, and inconsistency features are combined to detect fake news.

The major contributions of this paper are summarized to be three-fold:

• This paper presents a novel unified framework, TD-MMC, that can simultaneously capture the inconsistency and complementary of multi-modal information while also strengthening text-model importance.

• To reduce the fusion noise, the proposed model adopts a one-way cross-modal attention mechanism to selectively learn the social structure features and cross-model attention to acquire text-image interaction features and retain original textual information to detect fake news. This paper also introduces a new multi-modal loss to mitigate the impact of noise information due to data quality.

• Extensive experiments on two Chinese and English datasets have been conducted, and the results confirm TD-MMC's superiority over state-of-the-art approaches.

The rest of the paper is summarized as follows: Section 2 reviews the existing work related to fake news detection. Section 3 gives a new definition of multi-modal fake news. Section 4 describes TD-MMC in detail. In Section 5, a series of experiments are conducted, and analyze the experiment's results. Conclusions are provided in Section 6.

2 Related Work

Fake news detection has been a popular area in machine learning, which aims at recognizing fake news in different forms [14]. This section briefly reviews the relevant work from two perspectives: single-modal and multi-model fake news detection.

2.1 Single-Modal Fake News Detection

In single-modal tasks, early work mostly focused on mining text features using supervised models based on feature engineering [15]. For example, Castillo et al. [16] focused on posts, retweeted text posts, and built a topic classifier to identify the credibility of news. Liu et al. [17] designed an event detector for Twitter data to dynamically identify fast-moving news stories. The distinction between three types of news content (satirical, false, and real) was described by Horne et al. [18]. They implemented fake news detection by using support vector machines (SVM). Although the above methods have achieved some success, fake news detection using manual features is time-consuming and laborious.

Many scholars have tried to use deep learning methods to improve model performance. González et al. [19] designed a transformer encoder approach for English and Spanish to capture the contextual semantics of text words. Wang et al. [20] focused on global and local semantic relations and proposed a neural network model based on graphs, SemSeq4FD, for the early detection of fake news. Fake news based on images attracts more attention and spreads more widely than text. Shelke et al. [21] created a set of user features and trained a deep learning framework to improve the model's accuracy. Chen et al. [22] began at the participant level and mined the users' feature vectors to integrate into the fake news detection. Some prior research employed basic statistical aspects of the attached images to detect fake news, such as the number of illustrations [23]. For example, Qi et al. [24] tried to learn information in the frequency domain and pixel domain of images. They used the multi-branch CNN-RNN model to capture the features of fake news images from the physical and semantic levels to identify fake images. Propagation-based methods are often more effective than those based solely on text content. Xue et al. [25] introduced a novel time-based propagation framework to solve the problem that the propagation path of fake news comprises multiple dynamic graphs. This fake news detection approach effectively integrates three kinds of information: structure, content semantics, and time. On this basis, Yang et al. [26] proposed a PostCom2DR fake news detection model that first learns the relationship between posts and related comments using Graph Convolution Networks (GCN) and then uses the self-attention mechanism to remove irrelevant features. Inspired by this, this paper constructs heterogeneous graphs to obtain social graph representations using three types of information: user posts, original posts, and comments.

2.2 Multi-Modal Fake News Detection

It is difficult to learn textual and visual feature representations directly through the single-model framework. Therefore, multi-modal fake news detection has recently attracted more attention. Wang et al. [27] fused the text and image features extracted by Convolutional Neural Networks (CNN) and VGG-19 through the co-attention mechanism. Jin et al. [28] proposed a recursive neural network with an attention mechanism (att-RNN) to capture the cross-modal fusion of text and image information by attention mechanisms. However, simple multi-model fusion through series or concatenation operation not only inadequately learns cross-modal features but may also introduce fusion noise.

To overcome the above limitation, Dhawan et al. [29] proposed a multi-model fake news detection framework named GAME-ON, which realized granular interactions between text and images. But, the method only focuses on the complementary features between different models. Some scholars focused on distinguishing the consistency between text and images or text and external knowledge [30] to detect multi-model fake news. For example, Xiong et al. [31] proposed a multi-model fusion network (TRIMOON), which strengthens the dominant role of text patterns in news media through two image-text information inconsistency fusions but ignores the social graph structure information. Subsequently, some work has investigated how social network features can be integrated to help identify multi-model fake news. Zheng et al. [32] designed a multi-model feature-enhanced attention network (MFAN) that can integrate learning across modal representations of text, images, and social networks to perform fake news detection tasks. However, they ignore that there is a lot of disturbing information on social networks which can affect the model's performance. Song et al. [33] pointed out that the fusion of relevant information between different models while maintaining the unique properties of each mode can reduce fusion noise. Therefore, based on cross-modal attention residuals a multi-channel convolutional neural network (CARMN) was proposed. However, not all modal features are equally important, and images are proven to be only used as auxiliary information.

Table 1 shows the comparison of our study with related work. In summary, what makes our work different from other existing work is that: (1) Emphasizes the importance of textual information, and jointly uses text-image complementary features and text-image inconsistency features to detect fake news. (2) Utilizing cross-modal attention mechanisms to enhance text features through social networks and selectively learn useful features from social networks, And retaining the original text news to reduce fusion noise. (3) TD-MMC adds a new multi-model loss for better fusion.

Table 1: Comparison of related studies. Column notations: Textual Feature (TF), Visual Feature (VF), Social Graph Feature (SF), Text-Dominated feature (TD), Text-Image Complementary feature (T-I-C), Text-Image Inconsistency feature (T-I-I), Multi-modal Loss (ML)

	TF	VF	SF	TD	T-I-C	T-I-I	ML
[12]	✓	✓			✓		
[29]	✓	✓			✓		
[30]	✓	✓			✓	✓	
[27]	✓	✓	✓	✓	✓		
[31]	✓	✓		✓	✓	✓	
[32]	✓	✓	✓			✓	✓
TD-MMC	✓	✓	✓	✓	✓	✓	✓

3 Problem Formulation

Given M as a set of posts on social media, $M = \{T_1, T_2, T_3, \ldots, T_n\}$, where n is the number of the posts. Each post $T_i \in$ M consists of four parts: original texts, attached image, user history posts and comment contents, $T_i = \{t_i, v_i, u_i, r_i\}$, where t_i, v_i, u_i respectively indicate the text, image, user posts who published the post, and related comment contents. In there, $r_i = \{r_1, r_2, \ldots, r_m\}$ is a set of comments that correspond to the post T_i.

Following previous work [34], Fake news detection is generally modeled as a binary classification task. Each post T_i has its original true label $y_i \in \{0, 1\}$, where $y_i = 1$ means that the news is true and the opposite is fake. The multi-modal fake news detection task can be described as learning a detection function $f: \{t_i, v_i, u_i, r_i\} \rightarrow y_i$ to identify whether a given news story is false. Our goal is to train a model to simultaneously learn text, images, and social networks to effectively detect multi-modal fake news.

4 Methodology

4.1 Model Overview

TD-MMC consists of three modules: (1) Multi-model feature extraction module: text, image, and social network features are extracted by CNN, Residual Neural Network (ResNet), and Signed Graph Attention Networks (SiGATs). Next, those features are enhanced with multi-head self-attention thorough feature enhancement. (2) Multi-model feature fusion module: textural information is enhanced by a social graph structure, and then the cross-model feature of the enhanced text features and image features is obtained through the cross-modal attention mechanism. Additionally, the module captures the text-image similarity features and performs a hybrid fusion of text enhancement representations, cross-model fusion feature representations, and inconsistent representations. (3) The classification module produces classification labels. The overall structure of the proposed model is shown in Fig. 1.

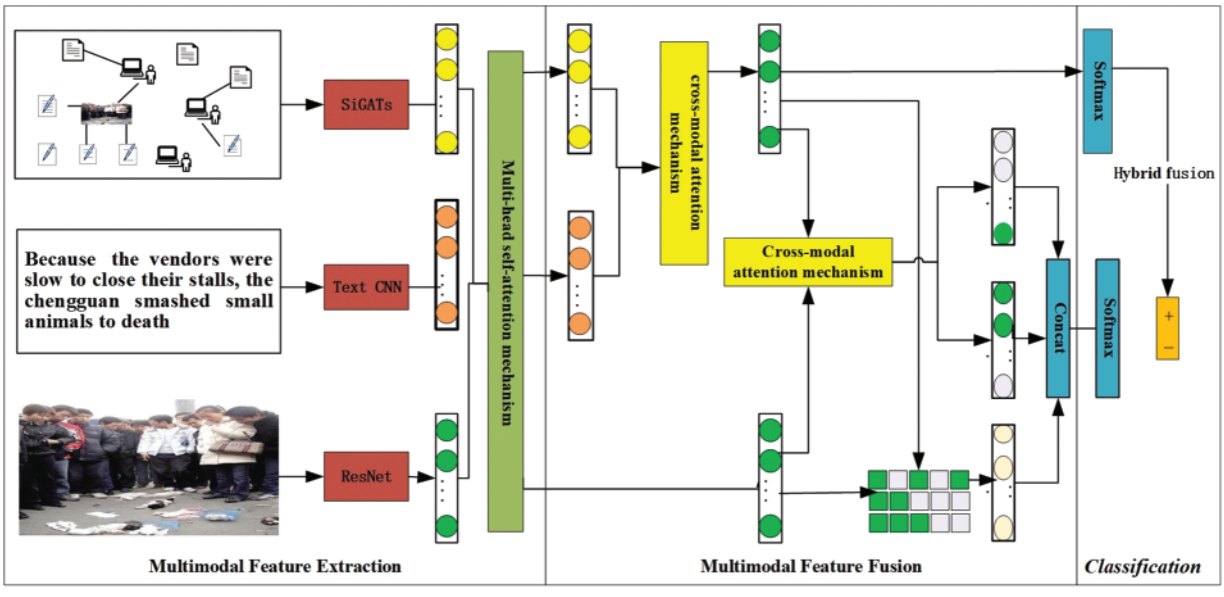

Figure 1: The proposed framework TD-MMC

4.2 Multi-Modal Feature Extraction Module

4.2.1 Textual Feature Extractor

In this section, CNN with pooling is used to extract textual content for each news article. Given a post T_i, the additional text content $t_i = \{w_1, w_2, w_3, \ldots, w_k\}$, where w_i represents the i-th word of the sentence t_i and the length of the sentence is k. The dimension of word embedding is d. First, CNN preprocesses the length of each post by setting the maximum length to L. The length of post texts smaller or larger than the length will be filled or truncated. The process can be described as:

$$t^i_{1:L} = \{w^i_1, w^i_1, \ldots w^i_L\} \tag{1}$$

After that, CNN obtains the feature map through convolutional layers and then performs a maximum pooling operation. Finally, the features obtained from each pooling layer are spliced to obtain the final text feature representation.

$$M_i = \sigma(W * X^i_{e:\, e+k-1} + b) \tag{2}$$

$$\hat{M} = max\{M\} = max\{[M_1, M_2, \ldots M_{L-K+1}]\} \tag{3}$$

$$h^t = concat(\hat{M}^i_{k=3} + \hat{M}^i_{k=4} + \hat{M}^i_{k=5}) \tag{4}$$

where, $*$ is the convolution operation, $b \in R$ is the bias term and α is an activation function. CNN uses *d/3* filters with varying receptive fields k $\in \{3, 4, 5\}$ to obtain the semantics from different granularities.

4.2.2 Image Feature Extractor

This paper employs the pre-trained model ResNet50 to extract image features [35]. For the attached visual content v_i of the post T_i, the process of extracting image features can be expressed as follows:

$$h^r = ResNet50(v_i) \tag{5}$$

The output of the last second layer of ResNet50 is extracted and represented it as $h^{r'}$. To make visual features have the same dimension as textual features for easy subsequent operations, ResNet50 passes $h^{r'}$ to a fully connected layer.

$$h^v = \sigma(W_v * h^{r'}) \tag{6}$$

where, W_v is the weight matrix and σ is the activation function.

4.2.3 Social Network Feature Extractor

Traditional Graph Attention Networks (GATs) make the correlation value between the query and the negative key very small after using the softmax function or even regard it as unimportant. However, the negative correlation may represent the opposite semantics. For example, there is a special node with neighboring nodes $n_i = \{-0.3, 0.7, -0.9, 0.2\}$. After the softmax function, it will become $\widehat{n_i} = \{0.08, 0.25, 0.19, 0.46\}$. It can be seen that the node corresponding to "-0.9" in the weight vector turns into "0.16", which means the node has the smallest contribution. However, "-0.9" may indicate that the two node vectors are in opposite directions. In the real world, a fake news spreader might buy some honest users as fans or post some comments to oppose the fake news, which creates negative correlations with the original news. This paper believes that the rational use of these features is beneficial to fake news detection.

SiGATs [36] are used to capture both positive and negative correlations in social networks. SiGATs first calculate the attention coefficient to determine the relevance of each node. The node embedding matrix includes three types of nodes: user posts, original posts, and comments as $Z = R^{(|V|\times d)}$, where d is the dimension size. The original post and comment nodes are represented by sentence vectors as the initial embedding and the embedding of user history post nodes is denoted by the average values as the initial user post embedding. In particular, SiGATs also add a "-softmax" operation, aiming to adopt the opposite value between the query and the key as input to the softmax function to amplify the negative correlation. The attention weights are as follows:

$$N_t^{'} = softmax(n_t) \tag{7}$$

$$\widetilde{N}_t^{'} = softmax(-n_t) \tag{8}$$

Then, concatenating the two vectors and obtaining the final node features through a fully connected layer. The process can be represented as:

$$N_t = concat(N_t^{'}, \widetilde{N}_t^{'}) \tag{9}$$

$$h^s = \sigma(W_N * N_t * X_j) \tag{10}$$

where, W_N is the fully connected weight matrix, σ is the activation function, and X_j is the feature matrix of N_t.

4.2.4 Feature Enhancement

To better capture the global semantic relations of each single-modal, the proposed model uses a multi-head self-attention mechanism to enhance text, image, and social structure features. It can process information from different locations in parallel by learning a variety of mappers through multiple linear mappings in each dimension of K, Q, and V [7]. Input the image feature h^v with three different initial weight distributions W_k, W_q and W_v. Firstly, $Q = W_q h^v$, $K = W_k h^v$, $V = W_v h^v$ are mapped through the matrix. The attention operation is repeatedly performed t times and the final result is obtained by splicing through the multiple self-attention mechanisms. The formula is shown below:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_K}}\right) V_t \tag{11}$$

$$head_i = Attention(QW_i^Q, KW_i^k, VW_i^v) \tag{12}$$

$$H_i = MultiHead(Q, K, V) = Concat(head_1, head_2, \ldots, head_m) * W_t \tag{13}$$

where, W_i^Q, W_i^k, W_i^v, W_t are learnable parameters and m is the number of attention heads.

Through the above process, this paper has obtained enhanced text, image, and social network features H_T, H_I, H_G. Then, the three enhanced features are converted into the same model feature space to better integrate textual and visual features, that is: $H_T^{'} = W_T H_T$, $H_I^{'} = W_I H_I$, $H_G^{'} = W_G H_G$ and W_T, W_I, W_G are learnable parameters.

4.3 Multi-Modal Feature Fusion Module

4.3.1 Cross-Modal Attention

In this section, a cross-modal attention network is used to capture the mutual information between different modalities. Unlike other attention mechanisms that only calculate self-attention on a single model, the cross-modal attention mechanism extends the calculation of attention to two modes, enabling more detailed features of the source model to be characterized [28] and effectively studying the complementary information between different modes.

Assuming that it needs to fuse text and image features, the multi-modal features $R = [H_T^{'}, H_I^{'}]^T$ are first generated through three fully connected layers: key feature matrix K_R, query feature matrix Q_R, and value feature matrix V_R. Then, the cross-modal attention network establishes the cross-modal relationship and calculates the similarity matrix between text and image by the scaled dot-product attention. The calculation process is as follows:

$$Attention(Q_R, K_R, V_R) = softmax\left(\frac{Q_R K_R^T}{\sqrt{d}}\right) V_R \tag{14}$$

To make the derivation simple and understandable, the softmax and scaling functions in the above equation are omitted and the equation can be extended as follows:

$$\begin{pmatrix}\hat{S}\\ \hat{I}\end{pmatrix} = Q_R K_R V_R = \begin{pmatrix}Q_{H_T^{'}}\\ Q_{H_I^{'}}\end{pmatrix}\begin{pmatrix}K_{H_T^{'}} & K_{H_I^{'}}\end{pmatrix}\begin{pmatrix}V_{H_T^{'}}\\ V_{H_I^{'}}\end{pmatrix} = \begin{pmatrix}Q_{H_T^{'}} K_{H_T^{'}}^T V_{H_T^{'}} + Q_{H_T^{'}} K_{H_I^{'}}^T V_{H_I^{'}}\\ Q_{H_I^{'}} K_{H_I^{'}}^T V_{H_I^{'}} + Q_{H_I^{'}} K_{H_T^{'}}^T V_{H_T^{'}}\end{pmatrix} \tag{15}$$

According to the above equations, the proposed model gets the text features fusing image features, $\hat{S} = Q_{H_T^{'}} K_{H_T^{'}}^T V_{H_T^{'}} + Q_{H_T^{'}} K_{H_I^{'}}^T V_{H_I^{'}}$, and the image features fusing text feature, $\hat{I} = Q_{H_I^{'}} K_{H_I^{'}}^T V_{H_I^{'}} + Q_{H_I^{'}} K_{H_T^{'}}^T V_{H_T^{'}}$.

In addition to post content, social network information is also beneficial for understanding the semantics of posts, but adding lots of irrelevant information affects the performance of the model, which is mostly ignored by existing multi-model fake news studies. This paper first utilizes the unidirectional cross-modal attention mechanism to selectively learn useful features in social networks to enhance text feature representation, as follows:

$$\hat{T} = Q_{H_T^{'}} K_{H_T^{'}}^T V_{H_T^{'}} + Q_{H_T^{'}} K_{H_G^{'}}^T V_{H_G^{'}} \tag{16}$$

After that, the interactive learning of information between enhanced text feature $\hat{T}$ and image feature $H_I^{'}$ is captured. According to the above formula, it can be obtained:

$$\widehat{T - I} = Q_{\hat{T}} K_{\hat{T}}^T V_{\hat{T}} + Q_{\hat{T}} K_{H_I^{'}}^T V_{H_I^{'}} \tag{17}$$

$$\widehat{I - T} = Q_{H_I^{'}} K_{H_I^{'}}^T V_{H_I^{'}} + Q_{H_I^{'}} K_{\hat{T}}^T V_{\hat{T}} \tag{18}$$

4.3.2 The Similarity Measurement

The similarity measurement module aims to assess the inconsistency of text and images by measuring the semantic similarity. Finding both semantically pertinent and non-manipulated images to support these non-factual stories is difficult. Therefore, the inconsistency feature between text and images can be easy to help identify fake news [6].

This section uses the cosine similarity to define the correlation between the enhanced text $\hat{T}$ and the visual information $H_T^{'}$. The formula is as follows:

$$h^s = \frac{\hat{T} * H_I^{'}}{\|\hat{T}\| * \|H_I^{'}\|} \tag{19}$$

where, $h^s \in [-1, 1]$. After that, the similarity is mapped between [0, 1] through the sigmoid activation.

$$H_S = sigmoid(h^s) \tag{20}$$

4.3.3 Hybrid Fusion

Until now, text enhancement features $\hat{T}$, text-image complementary features $\widehat{T-I}$, $\widehat{I-T}$, and text-image inconsistency feature H_S have been obtained. This section uses the hybrid fusion method to combine those features. It first concatenates $\hat{T}$, $\widehat{T-I}$, $\widehat{I-T}$ as X^i and feeds it into the fully connected layer to obtain the predicted probability y_m. Meanwhile, text-image inconsistent features H_S directly fed into the fully connected layer to obtain the predicted probability y_t. Finally, a late fusion between y_m and y_t is performed. The process is shown below:

$$X^i = concat(\hat{T}, \widehat{T-I}, \widehat{I-T})$$

$$y_m = softmax(w_c X^i + b)$$

$$y_t = softmax(w_s H_S + b)$$

$$\hat{y} = \alpha y_m + (1-\alpha) y_t \tag{21}$$

where, α is used to balance y_m and y_t, w_c and w_s are the weight matrix, and b is the bias term.

4.4 Classification Module

After the fused features are processed by the hybrid fusion, the classification results can be obtained. This paper built a new multi-model loss function to supervise training. Firstly, a binary cross-entropy defined loss function is adopted as fake news classification loss as follows:

$$L_{classify} = -y log(\hat{y}_i) - (1-y) log(1-\hat{y}_i) \tag{22}$$

where, y is the true label of the news and $\hat{y}_i$ is the predicted label.

In addition, the proposed model uses the MSE loss to measure the distance between matched image and text-embedded features to make them more tightly integrated and mismatching embedded features farther apart:

$$L_{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\hat{T} - H_I^{'})^2 \tag{23}$$

The final multi-model loss is defined as:

$$L = \varphi L_{classify} + (1-\varphi) L_{MSE} \tag{24}$$

where, φ is the equilibrium coefficient and $\varphi \in [0, 1]$.

5 Experiments

5.1 Datasets

This paper adopts two real-world datasets, covering English and Chinese, to conduct experiments. To learn multi-modal features better, this paper removes the data with missing modalities from the original data, and the final statistical results are shown in Table 2.

Table 2: The statistics of two datasets

	# Non-fake news	# Fake fake news	Users	Comments	Images
Weibo	877	590	985	4535	1467
PHEME	1428	590	894	7388	2018

Weibo. It is a Chinese dataset presented by Song et al. [37] and collected from Sina Weibo, which consists of 877 real news items and 590 fake news items. Each news item includes information on texts, attached images, and comments.

PHEME. This is an English dataset presented by Wei et al. [38], a selection of news pieces on the Twitter platform about five breaking news. Each post also contains textural, visual, and comment information.

5.2 Baselines

To verify the validity of the proposed model, this paper selects some state-of-the-art baselines for comparative analysis, which are divided into two groups: single-modal and multi-modal methods. The detailed introductions are below.

5.2.1 Single-Modal Methods

• Only-text: This method only uses the text content of the post to detect fake news. Text features are extracted from the CNN network. Input them into the self-attention mechanism to enhance text feature representation and then the fully connected layer for classification.

• Only-image: It only uses the image information in the post to detect fake news. The ResNet50 network is used to obtain visual features. The enhanced image information is then fed into the full connection and softmax layers for final prediction.

• Only-social graph: In this method, only social network information is retained. Social graph features are obtained through the SiGATs network and input into the full connection and softmax layers for fake news detection.

5.2.2 Multi-Model Methods

• EANN [27]: EANN is an adversarial neural network consisting of a multi-model feature extractor, a fake news detector, and an event discriminator that predicts the authenticity of a post while adding an event discriminator to predict the event label.

• Att-RNN [28]: Att-RNN learns a joint representation of text and social context and uses an attention mechanism to capture associations between visual features and joint text and social features to detect multi-model fake news.

• MFAN [33]: The multi-model feature-enhanced attention network (MFAN) fuses joint representation of multi-model data by co-attention mechanisms and considers cross-modal semantic alignment for multi-model fake news detection.

• MVAE [39]: MVAE uses a bimodal variational autoencoder that learns shared representations of text and images. It splices the separately obtained textual and visual feature representations and feeds them into a fully connected layer to form a shared representation, which is then used by a decoder to reconstruct the two modal features to detect fake news.

• SAFE [40]: It is also a multi-model fake news detection method. In this model, the representations of news textual and visual information, along with their relationships, are jointly learned and used to predict fake news.

• MMCN [41]: This multi-model fake news detection framework combines the multi-level semantics of text information with visual content to generate multi-level semantic features for fake news classification.

In short, EANN, MVAE, SAFE, and MMCN exploit textual and visual data. Att-RNN and MFAN also consider social graph features. EANN and MVAE fuse multi-model data through a simple fusion mechanism. MMCN and MFAN adopt cross-modal fusion methods to obtain interaction features between different modalities. In SAFE, image and text mismatch features are considered. None of them consider the importance of text modality as our proposed model does, and our work integrates the above studies.

5.3 Implementation Details

This paper randomly divides the above two datasets into training, validation, and testing sets with a ratio of 7:1:2. Following the previous work, the CNN model, the pre-trained ResNet50, and SiGATs are adopted to encode textural, visual, and social graph features. The parameters of the base model of feature encoding we use are frozen. The input size of the images is set to 224×224. For ease of computation, the proposed model adds a fully connected layer to make the image and text feature dimensions consistent in the multi-model feature fusion module, i.e., both image and text feature dimensions are 300. The final results are averaged over ten experiments for a fair comparison. This paper chooses the best parameter configuration based on the performance of the proposed model. The optimal values of φ to construct the loss function are 0.5 and 0.9 in the Weibo and PHEME datasets, respectively. Adam optimizer is used for optimization and sets the number of attention heads in the multi-head self-attention mechanism to 8. Other parameter settings involved in the model are shown in Table 3. The accuracy, precision, recall, and F1 score evaluation metrics in evaluating are used to evaluate the performance of the proposed model.

Table 3: Parameter settings of the proposed model

Parameters	Value
Batch size	64
Max epochs	20
α	0.5
Dropout	0.4

(Continued)

Table 3 (continued)	
Parameters	Value
Max length	50
Learning rate	0.02

5.4 Performance Comparison

In this section, the proposed model is compared with the above baselines based on single-model and multi-model to verify its effectiveness. Tables 4 and 5 show the experimental results on the PHEME and Weibo datasets. It can be drawn the following conclusions:

Table 4: Performance comparison of PHEME dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
Only-text	84.94	83.75	78.47	80.41
Only-image	68.31	69.49	68.31	68.81
Only-social graph	71.43	66.65	52.62	47.80
Att-RNN	85.00	80.09	82.40	82.90
EANN	77.13	71.39	70.07	79.87
SAFE	81.49	79.88	79.50	84.96
MVAE	75.62	73.49	72.25	70.34
MFAN	86.75	83.57	86.74	84.80
MMCA	87.20	86.25	85.00	85.55
TD-MMC	**89.61**	**90.10**	**84.37**	**86.57**

Table 5: Performance comparison of Weibo dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
Only-text	74.24	78.63	78.16	74.22
Only-image	72.88	71.56	71.13	71.31
Only-social graph	80.34	81.69	82.88	80.27
Att-RNN	7.20	78.40	77.25	76.85
EANN	80.96	80.19	79.68	79.87
SAFE	84.95	84.98	84.95	84.96
MVAE	71.76	70.52	70.21	70.32
MFAN	87.80	87.21	87.21	87.21
MMCA	87.90	87.95	87.95	87.65
TD-MMC	**90.17**	**89.68**	**89.76**	**89.72**

(1) The experimental results of the two datasets show that TD-MMC outperforms all baseline models in terms of accuracy, precision, recall, and F1 score. Compared with the optimal baseline, the accuracy, F1, ACC, and AUC of TD-MMC have increased whether on the PHEME or Weibo datasets, with accuracy improving by 2.39% and 2.27%, respectively. It proves that our model can effectively capture important multi-model clues ignored by the existing methods of detecting fake news.

(2) For single-modal models, it can be seen that the results of Only-social graph and Only-text are better than Only-image, proving that textual content provides richer cues and is more important than images [3]. Noticeably, the Only-social graph performs better than the Only-text on the Weibo dataset, but the opposite result is observed on the PHEME dataset. The reason is that social network information can provide the background content of posts but also bring in noisy information.

(3) When comparing single-model and multi-model methods, results show that the latter generally performs better than the former multi-model methods, indicating that multi-modal data can learn from each other to jointly detect fake news. However, EANN on the PHEME performs worse than the Only-text. The only social graph outperforms Att-RNN and MVAE on the Weibo dataset. These results show that simply fusing each modal feature only leads to accumulating irrelevant information and reduces the fake news classification performance.

(4) From the comparison of multi-model approaches, it can be observed that the SAFE method outperforms MVAE and EANN for multi-model methods but is less effective and robust than MFAN and MMCA. MVAE has the worst performance. Firstly, MVAE and EANN use only simple textual and visual information representation and sharing, which is insufficient for learning high-level shared semantics. EANN performs better than MVAE because adversarial training is used to reduce the impact of information loss. SAFE confirms that considering inconsistent information is beneficial for fake news detection. MMCA performs better among the baseline models. One possible reason is that the model learns the deep interaction features of text and image modes like MFAN but ignores the social graph structure. Our model outperforms Att-RNN and MFAN which all consider social graph features. The reason lies in using a one-way cross-modal attention mechanism to obtain enhanced text features by the social graph network, which only uses social graph network features as auxiliary features to supplement post text to reduce noise information on social networks.

5.5 Ablation Study

In this section, this paper designs ablation experiments to further validate the effectiveness of different key components in our model by discarding the related variables, which are defined as follows:

(1) W/o enhanced text feature: This variant removes the enhanced text feature in the fake news predictor module and only considers cross-modal fusion and inconsistency features.

(2) W/o inconsistency feature: This variant removes the similarity measurement module and does not consider the inconsistency features between text and image.

(3) Cross-modal fusion features: This variant only considers the cross-modal fusion feature of enhanced text and images for fake news detection.

(4) W/o multi-model loss: This variant only adopts the binary cross entropy definition loss function as the fake news classification loss.

(5) W/o social graph: This variant only uses the original text and image and removes social graph network information to detect fake news in this variant.

Tables 6 and 7 display the experimental results of our proposed model. It can be made the following observations:

Table 6: Results of ablation experiments on the PHEME dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
W/o enhanced text feature	87.01	83.99	85.64	84.73
W/o consistency module	88.05	86.47	84.04	85.12
Cross-modal module	85.71	82.28	82.64	82.73
W/o multi-model loss	83.64	80.36	79.88	80.15
W/o social graph	84.94	81.67	82.61	82.11
TD-MMC	**89.61**	**90.10**	**84.37**	**86.57**

Table 7: Results of ablation experiments on the Weibo dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
W/o enhanced text feature	87.12	86.31	87.56	86.73
W/o consistency module	88.14	87.35	88.71	87.78
Cross-modal module	87.46	86.99	88.78	87.22
W/o multi-modell loss	87.12	86.29	87.09	86.61
W/o social graph	87.46	86.21	86.37	86.29
TD-MMC	**90.17**	**89.68**	**89.76**	**89.72**

(1) In the two datasets, TD-MMC outperforms its five variants without the enhanced text feature, inconsistency feature, multi-model loss, social graph, and cross-modal fusion features, which indicates that removing those proposed components can reduce the performance of news classification.

(2) As seen from the results, the degree of importance: multi-model loss > enhanced text feature > consistency module, w/o multi-model loss is worse than other variants, proving that multi-model loss can benefit the models' performance. It aimed to refine the learned representations between each modality by continuously measuring the distance between image-text embedded features during training. This paper also finds that adding enhanced text features and a consistency module based on the cross-modal module can further improve accuracy. The reason may be that it mitigates the influence of noise information, which the cross-modal fusion component may generate. Additionally, by comparing the results of the TD-MMC and w/o social graph, it can be found that social networks can provide additional context for news and benefit the model's performance.

(3) The experimental results show that the model that removes multi-model loss performs much worse on the PHEME than Weibo. As we can see from Table 4, the accuracy of Only-text is higher at 16.18% and 13.51% higher, respectively, than Only-social graph and Only-image. However, there is little difference in the results between the three metrics on the Weibo dataset. This phenomenon may be caused by the difference in data quality between the two datasets. The PHEME dataset is characterized by low image quality and useless social network information. Therefore, multi-model supervised loss in the training process can also improve the model generalization ability.

5.6 Text-Model Importance Analysis

Previous work [42] has pointed out that original important information may be lost when the information of different modalities is fused. To solve this problem, some studies consider the original features of each model to minimize the loss of original information. However, it also reinforces the importance of image features. This paper believes the text model is the basis, and the image information is used as an auxiliary.

To validate our point, the following experiments are designed called TD-MMC∗ which simultaneously consider original image features. In the experiment, the Only-image method is also selected as the baseline for the comparison. The results of the experiment are shown in Fig. 2. It is obvious that TD-MMC achieves better performance than TD-MMC*, which proves that considering the original image features will reduce the performance of the model. This is because fake news images have more complex patterns on both a physical and semantic level, features that are harder to obtain with pre-trained models. The results also demonstrate that the Only-image method performs more poorly in fake news detection. Therefor, only paying attention to the auxiliary features of images in multi-model fake news detection is enough.

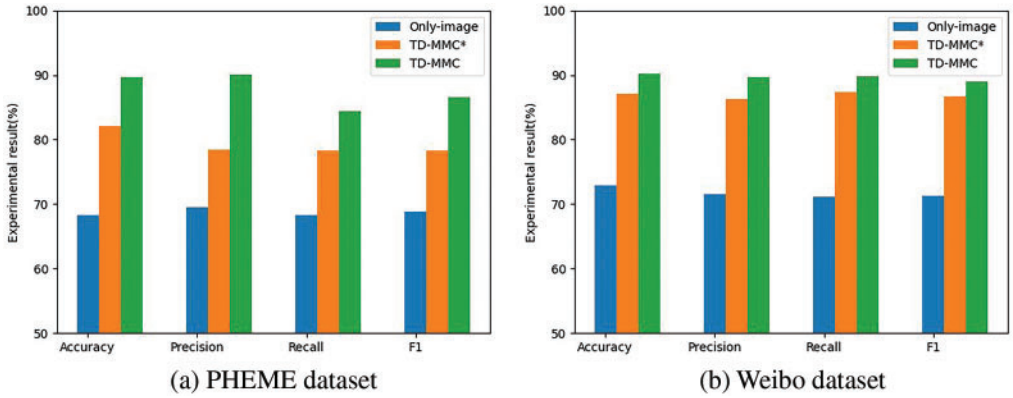

Figure 2: Text importance analysis

6 Conclusion

This paper proposes a novel multi-modal fake news detection method named TD-MMC, which effectively integrates multi-model data to generate more accurate, complementary, and comprehensive multi-model representations. The proposed model mainly consists of three modules: a multi-model feature extraction module, a multi-modal Feature Fusion module, and a classification module. First, the proposed model encodes the original text, image, and social graph network information. Then social network features are integrated into the text features representation to obtain enhanced text features as a supplement to the original text. To explore the features of text-image complementary, TD-MMC models the cross-modal attention mechanism to capture cross-fusion features of enhanced text and image features. It adopts the cosine similarity to obtain the text-image inconsistency features. Finally, fake news is classified by combining text enhancement features, text-image complementary features, and text-image inconsistency features. In addition, this paper establishes a new multi-model loss for training to improve the generalization ability of the model. Extensive experiments are conducted on real datasets. Compared with two baselines including single-modal and multi-model, the results show that TD-MMC is more effective than state-of-the-art baselines on PHEME and Weibo

datasets, with accuracy improving by 2.39% and 2.27%, respectively. The results also indicate that TD-MMC can effectively reduce the fusion noise to improve its performance.

Critical analysis and discussion: Although achieving success, TD-MMC has some limitations. Firstly, TD-MMC only extracts the text, image, and social graphs feature. In reality, there may also be other models' information such as video and audio. Secondly, the fusion technology of TD-MMC still has room for improvement. In future work, this paper will adopt a graph neural network to fuse more model information (text, image, video, audio, etc.), but also reduce the fusion noise. In addition, background knowledge from knowledge graphs can be considered integrated into multi-model fake news detection to improve model performance.

Acknowledgement: The authors would like to express their gratitude for the valuable feedback and suggestions provided by all the anonymous reviewers and the editorial team.

Funding Statement: This research was funded by the General Project of Philosophy and Social Science of Heilongjiang Province, Grant Number: 20SHB080.

Author Contributions: Conceptualization: Lifang Fu and Huanxin Peng; Data collection: Huanxin Peng; Analysis and interpretation of results: Lifang Fu and Huanxin Peng; Draft manuscript preparation: Huanxin Peng and Changjin Ma. Visualization: Yuhan Liu. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The datasets analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang and Y. Liu, "Combating fake news: A survey on identification and mitigation techniques," *ACM. Trans. Intel. Syst. Technol.*, vol. 10, no. 3, pp. 1–42, 2019. doi: 10.1145/3305260.

[2] F. Guo, A. Zhou, X. Zhang, X. Xu, and X. Liu, "Fighting rumors to fight COVID-19: Investigating rumor belief and sharing on social media during the pandemic," *Comput. Hum. Behav.*, vol. 139, pp. 107521, 2023. doi: 10.1016/j.chb.2022.107521.

[3] H. J. Alshahrani *et al.*, "Hunter prey optimization with hybrid deep learning for fake news detection on arabic corpus," *Comput. Mater. Contin.*, vol. 75, no. 2, pp. 4255–4272, 2023. doi: 10.32604/cmc.2023.034821.

[4] J. Ma, W. Gao, and K. F. Wong, "Rumor detection on twitter with tree-structured recursive neural networks," in *Proc. ACL-HLT*, Melbourne, Australia, 2018, pp. 1980–1989.

[5] P. Meel and D. K. Vishwakarma, "A temporal ensembling based semi-supervised ConvNet for the detection of fake news articles," *Expert Syst. Appl.*, vol. 177, pp. 115002, 2021. doi: 10.1016/j.eswa.2021.115002.

[6] H. Zhou, T. Ma, H. Rong, Y. Qian, Y. Tian and N. Al-Nabhan, "MDMN: Multi-task and domain adaptation based multi-modal network for early fake news detection," *Expert. Syst. Appl*, vol. 195, pp. 116517, 2022. doi: 10.1016/j.eswa.2022.116517.

[7] S. Singhal, R. Shah, T. Chakraborty, P. Kumaraguru, and S. I. Satoh, "Spotfake: A multi-modal framework for fake news detection," in *Proc. BigMM '19*, Singapore, 2019, pp. 39–47.

[8] Y. Wu, P. Zhan, Y. Zhang, L. Wang, and Z. Xu, "Multimodal fusion with co-attention networks for fake news detection," in *Proc. IJCNLP'21*, Bangkok, Thailand, 2021, pp. 2560–2569.

[9] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multi-modal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, 2017.
[10] L. Hu, Z. Zhao, X. Ge, X. Song, and L. Nie, "MMNet: Multi-modal fusion with mutual learning network for fake news detection," arXiv:2212.05699, 2022.
[11] P. Qi, J. Cao, X. Li, H. Liu, Q. Sheng and X. Mi, "Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues," in *Proc. MM '21*, New York, NY, USA, 2021, pp. 1212–1220.
[12] S. Qian, J. Wang, J. Hu, Q. Fang, and C. Xu, "Hierarchical multi-modal contextual attention network for fake news detection," in *Proc. SIGIR'21*, New York, USA, 2021, pp. 153–162.
[13] Q. Li, Q. Zhang, and L. Si, "Rumor detection by exploiting user credibility information, attention and multi-task learning," in *Proc. ANLP'19*, Florence, Italy, 2019, pp. 1173–1179.
[14] L. Fu, H. Peng, and S. Liu, "KG-MFEND: An efficient knowledge graph-based model for multi-domain fake news detection," *J. Supercomput.*, vol. 79, pp. 18417–18444, 2023. doi: 10.1007/s11227-023-05381-2.
[15] 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.
[16] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proc.WWW'11*, New York, USA, 2011, pp. 675–684.
[17] X. Liu, A. Nourbakhsh, Q. Li, R. Fang, and S. Shah, "Real-time rumor debunking on twitter," in *Proc. CIKM'15*, New York, USA, 2015, pp. 1867–1870.
[18] B. Horne and S. Adali, "This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news," in *Proc. ICWSM'17*, Montreal, Canada, 2017, pp. 759–766.
[19] J. Á. González, L. F. Hurtado, and F. Pla, "Transformer based contextualization of pre-trained word embeddings for irony detection in Twitter," *Inf. Process. Manag.*, vol. 57, no. 4, pp. 102262, 2020. doi: 10.1016/j.ipm.2020.102262.
[20] Y. Wang, L. Wang, Y. Yang, and T. Lian, "SemSeq4FD: Integrating global semantic relationship and local sequential order to enhance text representation for fake news detection," *Expert. Syst. Appl.*, vol. 166, pp. 114090, 2021. doi: 10.1016/j.eswa.2020.114090.
[21] S. Shelke and V. Attar, "Fake news detection in social network based on user, content and lexical features," *Multimed. Tools. Appl.*, vol. 81, no. 12, pp. 17347–17368, 2022. doi: 10.1007/s11042-022-12761-y.
[22] X. Chen, F. Zhou, F. Zhang, and M. Bonsangue, "Catch me if you can: A participant-level fake news detection framework via fine-grained user representation learning," *Inf. Process. Manag.*, vol. 58, no. 5, pp. 102678, 2021. doi: 10.1016/j.ipm.2021.102678.
[23] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel visual and statistical image features for microblogs news verification," *IEEE Trans. Multimed.*, vol. 19, no. 3, pp. 598–608, 2016. doi: 10.1109/TMM.2016.2617078.
[24] P. Qi, J. Cao, T. Yang, J. Guo, and J. Li, "Exploiting multi-domain visual information for fake news detection," in *Proc. ICDM'19*, Beijing, China, 2019, pp. 518–527.
[25] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi and L. Wei, "Detecting fake news by exploring the consistency of multi-model data," *Inform. Process. Manag.*, vol. 58, no. 5, pp. 102610, 2021. doi: 10.1016/j.ipm.2021.102610.
[26] Y. Yang, Y. Wang, L. Wang, and J. Meng, "PostCom2DR: Utilizing information from post and comments to detect fake news," *Expert. Syst. Appl.*, vol. 189, pp. 116071, 2022. doi: 10.1016/j.eswa.2021.116071.
[27] Y. Wang *et al.*, "EANN: Event adversarial neural networks for multi-modal fake news detection," in *Proc. SIGIR'24*, New York, USA, 2018, pp. 849–857.
[28] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proc. MM '17*, New York, NY, USA, 2017, pp. 795–816.
[29] M. Dhawan, S. Sharma, A. Kadam, R. Sharma, and P. Kumaraguru, "GAME-ON: Graph attention network based multimodal fusion for fake news detection," arXiv:2202.12478, 2022.

[30] M. Sun, X. Zhang, J. Ma, S. Xie, Y. Liu and S. Y. Philip, "Inconsistent matters: A knowledge-guided dual-consistency network for multi-modal rumor detection," in *Proc. EMNLP*, Punta Cana, Dominican Republic, 2021, pp. 1412–1423.
[31] S. Xiong, G. Zhang, V. Batra, L. Xi, L. Shi and L. Liu, "TRIMOON: Two-round inconsistency-based multi-modal fusion network for fake news detection," *Inf. Fusion*, vol. 93, pp. 150–158, 2023. doi: 10.1016/j.inffus.2022.12.016.
[32] J. Zheng, X. Zhang, S. Guo, Q. Wang, W. Zang and Y. Zhang, "MFAN: Multi-modal feature-enhanced attention networks for rumor detection," in *Proc. IJCAI'22*, Barcelona, Spain, 2022, pp. 2413–2419.
[33] C. Song, C. Yang, H. Chen, C. Tu, Z. Liu and M. Sun, "CED: Credible early detection of social media fake news," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 8, pp. 3035–3047, 2019. doi: 10.1109/TKDE.2019.2961675.
[34] A. Kishwar and A. Zafar, "Fake news detection on Pakistani news using machine learning and deep learning," *Expert. Syst. Appl.*, vol. 211, pp. 118558, 2023. doi: 10.1016/j.eswa.2022.118558.
[35] Z. Wu, C. Shen, and A. van den Hengel, "Wider or deeper: Revisiting the ResNet model for visual recognition," *Pattern Recognit.*, vol. 90, pp. 119–133, 2019. doi: 10.1016/j.patcog.2019.01.006.
[36] J. Ma, W. Gao, Z. Wei, Y. Lu, and K. F. Wong, "Detect rumors using time series of social context information on microblogging websites," in *Proc. CIKM'15*, New York, NY, USA, 2015, pp. 1751–1754.
[37] C. Song, N. Ning, Y. Zhang, and B. Wu, "A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks," *Inf. Process. Manag.*, vol. 58, no. 1, pp. 102437, 2021. doi: 10.1016/j.ipm.2020.102437.
[38] L. Wei, D. Hu, W. Zhou, Z. Yue, and S. Hu, "Towards propagation uncertainty: Edge-enhanced Bayesian graph convolutional networks for rumor detection," arXiv:2107.11934, 2021.
[39] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "MVAE: Multimodal variational autoencoder for fake news detection," in *Proc. WWW'19*, New York, NY, USA, 2019, pp. 2915–2921.
[40] X. Zhou, J. Wu, and R. Zafarani, "Fake news: Fundamental theories, detection strategies and challenges," in *Proc. WSDM '19*, New York, USA, 2020, pp. 354–367.
[41] I. Segura-Bedmar and S. Alonso-Bartolome, "Multi-modal fake news detection," *Inf.*, vol. 13, no. 6, pp. 284, 2022. doi: 10.3390/info13060284.
[42] D. Kumar Dixit, A. Bhagat, and D. Dangi, "Fake news classification using a fuzzy convolutional recurrent neural network," *Comput. Mater. Contin.*, vol. 71, no. 3, pp. 5733–5750, 2022. doi: 10.32604/cmc.2022.023628.