Complex & Intelligent Systems (2025) 11:303
https://doi.org/10.1007/s40747-025-01919-4

ORIGINAL ARTICLE

Contrastive learning of cross-modal information enhancement for multimodal fake news detection

Weijie Chen[1] · Fei Cai[1] · Yupu Guo[1] · Zhiqiang Pan[1] · Wanyu Chen[2] · Yijia Zhang[2]

Received: 5 June 2024 / Accepted: 28 April 2025 / Published online: 22 May 2025

Abstract

With the rapid development of the Internet, the existence of fake news and its rapid spread has brought many negative effects to the society. Consequently, the fake news detection task has become increasingly important over the past few years. Existing methods are predominantly unimodal methods or the multimodal representation of unimodal fusion for fake news detection. However, the large number of model parameters and the interference of noisy data increase the risk of overfitting. Thus, we construct an information enhancement and contrast learning framework by introducing Improved Low-rank Multimodal Fusion approach for Fake News Detection (ILMF-FND), which aims to reduce the noise interference and achieve efficient fusion of multimodal feature vectors with fewer parameters. In detail, an encoder extracts the feature vectors of text and images, which are subsequently refined using the Multi-gate Mixture-of-Experts. The refined features are mapped into the same space for semanteme sharing. Then, a cross-modal fusion is performed, resulting in that an efficient and highly precision fusion of text and image features is done with fewer parameters. Besides, we design an adaptive mechanism that can adjust the weights of the final components according to the modal fitness before inputting them into the classifier to achieve the best detection results in the current state. We evaluate the performance of ILMF-FND and the competitive baselines on two public datasets, i.e., Twitter and Weibo. The results indicate that our ILMF-FND greatly minimizes the number of parameters while outperforming the best baseline in terms of accuracy by 0.2% and 1.1% on the Weibo and Twitter datasets, respectively.

Keywords Fake news detection · Multimodal fusion · Contrastive learning

Introduction

Over the span of years, the pervasive application of social media platforms has facilitated users to easily share their daily experiences and thoughts through posts. However, it also creates conditions for the spread of fake news [1–3]. As a result, the fake news detection leads to a wide public concern of many researchers. So far, most of the research about fake news detection has paid too much attention to unimodal text, ranging from manual detection methods [4, 5], machine learning methods [6–8] to deep learning methods [9–11], which receives an over-focus on the extraction of semantic features from text for detecting fake news. However, most news in social media contains multimodal information, and the interweaving of the information reduces the accuracy of unimodal detection methods [12]. Therefore, there is an urgent need for a technique that can simultaneously take into account the information of different modalities to improve the current situation of limited unimodal detection.

The concept about cross-modal fusion emerges as a promising approach to address these issues by integrating the information features from multimodal to enhance the detection performance [13]. As shown in Fig. 1a, b, juxtaposing textual content with accompanying images can reveal the different signal of potential fake news. This emphasizes the importance of modal collaboration in enhancing the detection accuracy. However, indiscriminate cross-modal fusion may introduce noise and compromise detection precision [11, 13]. This requires us to achieve a refined cross-modal fusion but does not imply an over-reliance on it. Meanwhile, a key common limitation of existing methods is that they completely rely on the multimodal features while ignoring their intrin-

✉ Fei Cai
caifei08@nudt.edu.cn

1 Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha, Hunan, China

2 College of Electronic Engineering, National University of Defense Technology, Hefei, Anhui, China

(a) A young scientist is thinking down, putting a lot of effort and sweat into the pursuit of scientific breakthroughs.

(b) You left in peace, left me in pieces.

(c) Smoking is good for your health and can reduce your risk of cancer.

(d) When his wife divorced him, she divided all the property in half.

Fig. 1 Examples of multimodal news analysis. **a** The image does not add substantial information while the text content mismatches with image. **b** The text and image both reflect sadness. **c** The text goes against common sense so it is obviously fake. **d** The doctored image suggests that it is probably fake news

sic connection with the unimodal features, which leads to an increase in algorithm complexity and unsatisfactory detection results. For example, the text in Fig. 1c or the image in Fig. 1d alone can indicate that they are fake news because they go against the common sense. In addition, many approaches usually involve a large number of parameters in cross-modal fusion, which leads to high algorithmic complexity and inefficient detection. For this purpose, a Low-rank Multimodal Fusion is proposed [14], which adopts tensor decomposition to reduce the number of parameters to some extent. Nevertheless, this approach is not effective in reducing the large number of parameters in cross-modal fusion, and also exposes the model to the risk of overfitting. To address these defects, we propose an integrated approach that brings unimodal and multimodal features together for multimodal fake news detection. Meanwhile, we also introduce the modal fitness which is used to measure the modal correlations. This requires not only a careful consideration of modal fitness, but also an adaptive combination of unimodal and multimodal features to optimize the fake news classification based on the modal fitness.

In this paper, we first give a definition of modal fitness (MF) from the perspective of information theory and use the distribution of unimodal features to make modal fitness quantifiable. Then, we propose an Improved Low-rank Multimodal Fusion approach for Fake News Detection (ILMF-FND), which is composed of three modules. 1) An *Information Extraction Alignment Module*, which extracts the features of text and images in multimodal news to form the feature vectors by specific modal feature extractors. Then, the Multi-gate Mixture-of-Experts (MMoE) is utilized to achieve information enhancement and refinement of the feature vectors of each modal. It also maps the heterogeneous unimodal features into the shared semantic space. 2) A *Cross-modal Fusion Module*, which performs the cross-modal feature fusion by the Improved Low-rank Multimodal Fusion (ILMF) to obtain the correlation between modalities. 3) A *Model Adaption Module*, which quantifies the modal fitness by the Kullback-Leibler(KL) divergence between the unimodal features and realizes the redistribution of weights. Subsequently, ILMF-FND adaptively adjusts the weight of unimodal features according to the modal fitness. The results of comprehensive experiments show that ILMF-FND outperforms the original baseline model in terms of accuracy on two widely used datasets Twitter and Weibo. To sum up, our contributions in this paper can be listed as follows:

(1) We propose a novel concept of modal fitness, which is mainly used to measure the difference between modalities and quantify the modal fitness using the relationship of unimodal distributions.
(2) We introduce a novel fake news detection model ILMF-FND, which decomposes the weights into low-rank factors by ILMF. It also realizes the effective fusion of cross-modal features while reducing the number of parameters, and makes the final optimization of the model by combining with the measure of modal fitness.
(3) We conduct series of experiments on two widely used datasets Twitter and Weibo to evaluate our model. The consequences indicate that ILMF-FND performs better with fewer parameters for the task of fake news detection than the original baseline model.

We organize the rest of the paper as follows. "Related work" section describes some important work about fake news detection. "Preparatory concept" section describes the preparatory knowledge required in the framework of the detection method. "Approach" section introduces our proposed multimodal fake news detection method in detail. "Experiments" section describes the experimental procedure and discusses the results from different perspectives. Finally, "Conclusion" section is the conclusion of our work.

Related work

We discuss the related work from two perspectives, i.e., initial unimodal approaches ("Unimodal approach" section) and present multimodal approaches ("Multimodal approach" section).

Unimodal approach

A large number of works so far have paid much attention to unimodal. This situation is inextricably associated with the fact that early social media tended to be presented to the public in a single form [15]. Specifically, unimodal feature information is used to achieve fake news detection: analyze research using textual content [16, 17] or identify fake news by parsing images [18]. Most of the previous methods start from the shallow syntactic analysis of the text, such as the encoding method based on word class labels [19]. In detail, inspired by Generative Adversarial Networks (GANs), Ma et al. [20] propose a GAN-style approach that aims to make classifiers learn stronge rumor representations by generators creating the rumors in the training set.

This is actually a passive learning model that utilizes more challenging fake news to make the classifier focus on learning the low-frequency yet discriminative patterns.

However, various natural language processing tasks are gradually replaced by large-scale pre-trained models with the appearance of transformer models.

For instance, Sun and Chen [21] combine lite bidirectional encoder representations from transformers (ALBERT) and long short-term memory (LSTM) to dig into the deeper features of the text.

Moreover, Mayank et al. [22] obtain entity and relational features by performing named entity recognition on news headlines and encode them in combination with tensor decomposition models.

However, these methods ignore the cross-modal features, which may pull down the overall performance of multimodal news. Only considering unimodal can lead to inadequate utilization of information and limited accuracy [23]. Furthermore, it is challenging to acquire a model that exhibits a strong generalization capabilities because of the limited dataset size.

Multimodal approach

In addition to the above unimodal approach, scholars have also proposed many multimodal fake news detection methods to improve accuracy in recent years [11, 24–26]. They attempt to take advantage of the information mentioned in news more effectively through a multimodal approach [24]. Earlier work simply fuse modalities [18] and jointly consider text and images to accomplish the task of fake news detection [27]. To further explore the advantages of multimodal methods, Jin et al. [28] investigate and illustrate the application of multimodal machine learning in fake news detection, pointing out possible future research directions. Based on this, Zhang et al. [27] propose a novel Multimodal Knowledge-aware Event Memory Network (MKEMN) approach to improve rumor detection by using external knowledge to supplement the semantic representation of news text. Meanwhile, some of scholars argue that characteristics of different topics or events and information within certain specific domains should also be taken into account [29, 30].

However, numerous studies have focused on the refinement of modal features [25] and the innovation of cross-modal fusion methods [24, 26] in recent years. In particular, Chen et al. [24] propose an ambiguity-aware multimodal fake news detection method CAFE, which well captures the correlation between modalities and solve the misclassification problem caused by the inconsistency of different modalities. Inspired by Chen [24], Wang et al. [26] then introduce a cross-modal contrastive learning framework, which achieves an accurate text-image alignment using an attention mechanism with an attention-guided module. Besides, they give an explainability for cross-modal correlation. All in all, these ideas and models enable a certain degree of improvement in accuracy [31]. Nevertheless, Qi et al. [25] feel that refining

the feature vectors before cross-modal fusion would lead to better fusion and higher accuracy. They then use the refined feature representations to roughly predict the reality of the news and reweigh the representation of each modal by the prediction score to achieve better prediction.

Previous studies in the realm of multimodal fake news detection achieve significant advancements, yet the integration of refined unimodal features with consistent cross-modal representations remains a formidable challenge. Furthermore, a large number of parameters are involved in these researches, which may lead to poor utilization of information and thus reduce the model efficiency. On the contrary, our work well integrates unimodal features with cross-modal features and achieves a reduction in the number of parameters while improving detection efficiency, which decreases the complexity of our algorithm to some extent.

Preparatory concept

In this section, in order to facilitate the subsequent description, we elaborate on the key issue that needs to be used in our paper, i.e., cross-modal correlations. Given a dataset $\mathcal{D}$, which is composed of multimodal information and can be represented as $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$, where $\mathcal{X}$ is the set of modalities in posts and $\mathcal{Y}$ is the set of labels in $\mathcal{X}$. For example, a sample $\mathcal{A}$ will be denoted as $\{x, y\} = \left\{\left\{x^i\right\}_n, y\right\} \in \mathcal{D}$,where x^i is the i-th modality in $\mathcal{A}$. The modality can be text, image, etc. in a multimodal problem, and y is the label of $\mathcal{A}$.

Multimodal fake news detection is essentially a simple binary classification task to identify whether the posts are true or false. Specifically, it aims to map posts to the most likely labels by learning their rich feature inputs x_i, including unimodal features, multimodal features and modal fitness. Compared with the previous text classification problems, its uniqueness is the modal correlation, i.e., the modal fitness. It is a non-negligible factor affecting the classification effect. In order to understand this problem more deeply, we give the definition of modal fitness and the related description in our paper.

Definition 1 (modal fitness). For a given news sample $\{x, y\} \in \mathcal{D}$, the modal fitness $\varepsilon_{i,j}$ between the i-th modality and the j-th modality is defined as a distance between two unimodal modalities, i.e., $dis(i, j)$.

In this paper, only the modal fitness between text and image modalities is considered, and the distance between these two modalities is represented quantitatively by transforming the fitness into a unimodal feature with the KL divergence [24] as *dis(text, image)*.

With respect to the examples given in the first section, the existence of modal fitness is a key factor in the ability of our method to adaptively adjust the relationship between the unimodal and cross-modal fusion features. In other words, the modal fitness is a measure that quantifies the correlation between modalities. If the fitness between two modalities is low, it means that the information difference between them is large, and it is necessary to make a comprehensive judgment through modal information fusion. On the contrary, if the modal fitness between two modalities is high, it is possible to realize the accurate classification through the information of a single modality. Therefore, the proposal and quantification of modal fitness provide a better measure of the difference in information between two modalities and compensates for the lack of unimodal information.

In our work, we tackle and utilize the cross-modal correlations to achieve better fake news detection results.

Briefly speaking, two general modalities are included in a given dataset, which are text and image.

And our purpose is to compute the modal fitness between the above two modalities to reflect the cross-modal correlations.

Specifically, modal fitness can be measured by KL divergence, which is a quantitative measure of the distance between unimodal distributions.

Approach

In this section, the ILMF-FND framework proposed is composed of three key components:

1) Information extraction alignment ("Information extraction alignment module" section), including a unimodal feature information extractor and cross-modal information alignment.
2) Cross-modal fusion ("Cross-modal fusion module" section), which is obtained by converting the aligned vector compression into a low-rank tensor.
3) Modal adaptation ("Model adaption module" section), the weight allocation of unimodal features and cross-modal fusion features is re-realized according to the modal adaptation between text and image, so as to improve the accuracy of subsequent classification.

Finally, to better understand our model, we show the flow of ILMF-FND in "ILMF-FND algorithm" section by an algorithm. We detail the framework of ILMF-FND in Fig. 2.

Information extraction alignment module

Drawing on previous ideas, we transform the data into vector representations, i.e., the text and image information in each posts is transformed into its corresponding feature vector by the means of a specific modal encoder [32] to improve the machine's readability and manipulability of the data. Consid-

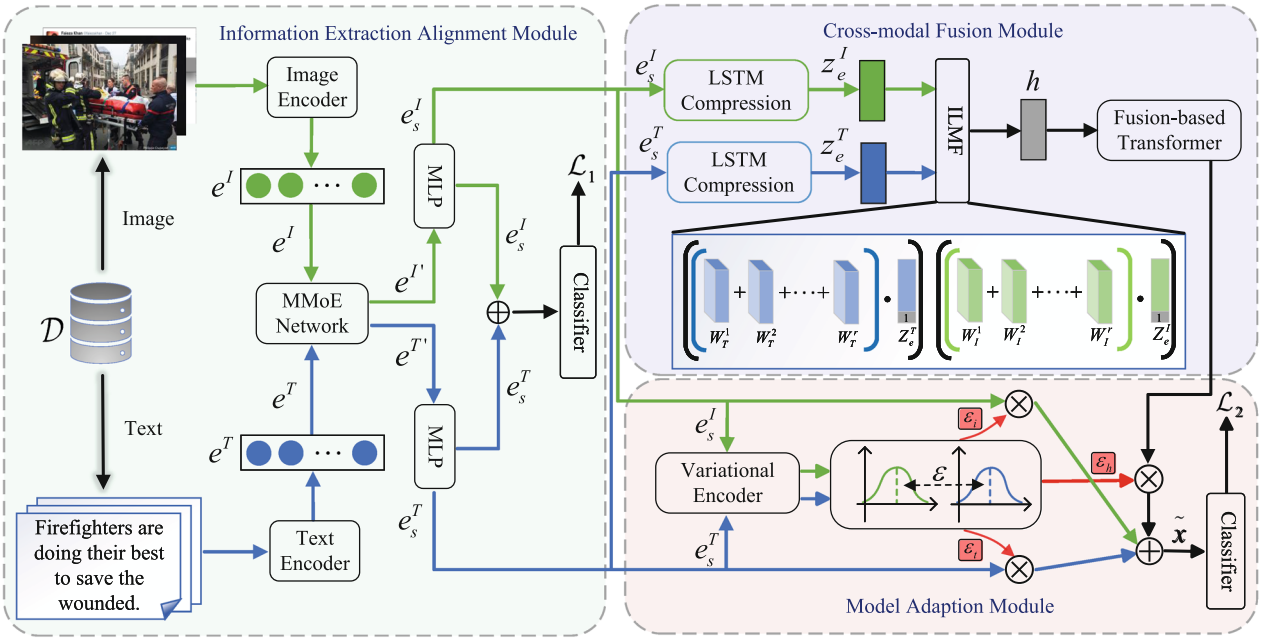


Fig. 2 Architecture diagram of the proposed ILMF-FND method. Unimodal features are first extracted from the multimodal news. Next, Multi-gate Mixture-of-Expert network refines these features, which are then fused in the cross-modal fusion module. The model adaption module can adaptively realize the allocation of weights to each modality according to the modal fitness, where each unimodal weight is ε_i and ε_t respectively. The multimodal weight is ε_h. When the modal fitness is high, the unimodal weights will adaptively become larger, making it better to reflect the role of unimodal features in fake news detection and thus improve detection efficiency. On the contrary, cross-modal fusion features take the dominant role. Finally, the $\oplus$ and $\otimes$ symbols represent the addition and multiplication operation, respectively

ering the advanced capabilities of BERT [33] and ResNet-34 [34] in understanding both text and image features, they are adopted as the primary tools for feature extraction in this module.

For encoding the text, given a text x^T, which is a collection of words. It is fed into a pre-trained BERT model [33], which is then passed through a fully connected layer to obtain the corresponding text embedding vector $e^T \in \mathbb{R}^{d_1}$. For the image encoder, given an image x^I, we apply ResNet-34 [34] to extract the meaningful parts of the image and characterize them. The results obtained from the characterization are also sent through a fully connected layer to capture the image embedding vector $e^I \in \mathbb{R}^{d_2}$.

After extracting the modal features in the first step, the embedding vectors e^T and e^I of text and image may have a large semantic gap. So it is necessary to map the respective unimodal embedding vectors into the shared semantic space to facilitate the subsequent cross-modal feature fusion. We thus introduce an auxiliary task about correlation learning to better realize the cross-modal alignment work. In other words, we use a simple encoder to build different visual and linguistic semantic levels to form a cross-modal contrast learning framework.

This auxiliary correlation learning task is just a simple binary classification task, i.e., given the features of text and images of a particular news report to predict whether the text and image pair is positively correlated (Real) or negatively correlated (Fake). First, we construct a new dataset $\mathcal{D}_1 = [\mathcal{D}_{pos}, \mathcal{D}_{neg}]$ based on the original dataset $\mathcal{D} = [\mathcal{D}_{train}, \mathcal{D}_{test}]$, where $\mathcal{D}_{pos}$ represents the positive set and $\mathcal{D}_{neg}$ represents the negative set while $\mathcal{D}_{trian}$ represents the training set and $\mathcal{D}_{test}$ represents the test set. If the text and image embeddings are from the same real news, then they belong to $\mathcal{D}_{pos}$ and are labeled $y_1 = 1$. Conversely, apart from the situation above, they are classified as $\mathcal{D}_{neg}$ and labeled $y_1 = 0$.

To enhance the alignment of multimodal features, we develop a contrastive learning model. The model consists of Multi-gate Mixture-of-Experts (MMoE) [35], modality-specific Multi-Layer Perceptron and modality sharing layer. Initially, we use MMoE to extract meaningful text and image feature parts. We then feed the resulting meaningful features into two parallel Multi-Layer Perceptrons (MLPs) to obtain the modality features. Subsequently, the resulting shared semantic embeddings are input to the average pooling layer and processed through a binary classifier constructed with a fully connected layer.

We introduce the cosine loss with margin m to train the positive and negative pairwise sets, the mathematical repre-

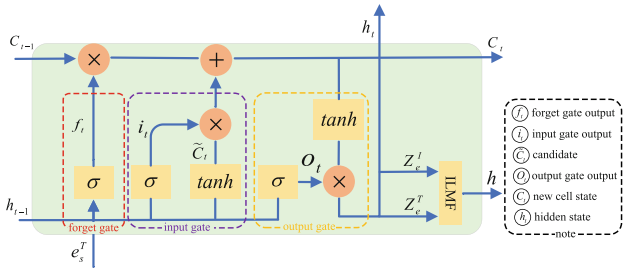


Fig. 3 Refined figure of the cross-modal fusion module, which is a continuous link graph, only one of them is shown here. The inputs are feature vectors of text and images after sharing semantics, and σ represents the sigmoid function. Only the e_s^T is shown here as an example

sentation of which is shown below:

$$\mathcal{L}_1 = \begin{cases} 1 - \cos(e_s^T, e_s^I), & if\, y_1 = 1 \\ \max(0, \cos(e_s^T, e_s^I) - m), & if\, y_1 = 0 \end{cases}, \tag{1}$$

where $cos\,(\cdot)$ is the cosine similarity after normalization, and m is the margin value, which is set to 0.2 based on the experience of previous studies [26]. The purpose of this section is to make positive text-image pairs close to each other and negative text-image pairs far away from each other. Meanwhile, semantic regularization can automatically map heterogeneous multimodal embeddings into a shared semantic space by back-propagated gradients.

Finally the obtained unimodal representations e_s^T, e_s^I of the semantically aligned text and image are used as inputs to *Cross-modal Fusion Module* for feature fusion and *Modal Adaption Module* for computing the modal fitness.

Cross-modal fusion module

In order to facilitate an effective modality-specific fusion, we utilize LSTM [36] and ILMF [14] to implement the feature fusion. Inspired by Liu et al. [37] in the field of sentiment analysis, a low-rank tensor is used to realize multimodal feature fusion in this paper while discarding the triple Cartesian product previously embedded from a specific modality [38].

In detail, to avoid the long dependency problem, we first use LSTM to compress the time series information of each modality and extract the hidden state context vectors for modality-specific fusion. Figure 3 gives a refinement of the cross-modal fusion process and finally gets the cross-modal fusion features h.

First, the sigmoid layer of the forget gate decides what information we want to discard from the input features. It outputs a number value between 0 and 1 for the previous cell state C_{t-1} according to h_{t-1} (previous link output) and e_t (current input), with 1 retaining the information in its entirety and 0 deleting it completely.

We then calculate i_t and $\widetilde{C}_t$ by the activation functions $sigmoid$ and $tanh$, respectively, where the $sigmoid$ function determines the output information and the $tanh$ function processes the state of the memory cell to prepare the output.

Next, we multiply the previous state value by f_t to express the part expected to be forgotten, after that we add $i_t * \widetilde{C}_t$ to get the new cell state C_t. Finally, we then use a sigmoid layer to determine the final input vector and normalize it between –1 and 1 to get the output h_t.

$$\begin{cases} f_t = \sigma\left(W_1 \cdot \left[h_{t-1}, e_s^T\right] + b_1\right) \\ i_t = \sigma\left(W_2 \cdot \left[h_{t-1}, e_s^T\right] + b_2\right) \\ \widetilde{C}_t = tanh\left(W_3 \cdot \left[h_{t-1}, e_s^T\right] + b_3\right), \\ C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \\ o_t = \sigma\left(W_4\left[h_{t-1}, e_s^T\right] + b_4\right) \\ h_t = Z_e^T = o_t * tanh\,(C_t) \end{cases} \tag{2}$$

where W_1, W_2, W_3 and W_4 are all learnable weight matrices and b_1, b_2, b_3 and b_4 are bias vectors. Similarly, the computation for the e_s^I in question is conducted using the same methodology as before.

Before performing the outer product, the unimodal tensor sequence is added by 1. This corresponds to the representation of tensor, which can capture the interaction between unimodal and multimodal

$$Z = \begin{bmatrix} Z_e^T \\ 1 \end{bmatrix} \otimes \begin{bmatrix} Z_e^I \\ 1 \end{bmatrix}. \tag{3}$$

Finally, the compressed representation h is generated through the batch matrix multiplications of the low-rank modality-specific elements and the appended modality portrayals:

$$h = \left(\sum_{i=1}^{r} \bigotimes_{j=1}^{J} W_j^{(i)}\right) \cdot Z = \bigwedge_{j=1}^{J}\left[\sum_{i=1}^{r} W_j^{(i)} \cdot z_j\right], \tag{4}$$

where $W_j^{(i)}$ is the low-rank factor of modal j and $\Lambda_{j=1}^{J}$ is denoted as the elemental product of a series of tensors. Here, J indicates the number of modalities, i.e., $J = 2$ in this paper.

Model adaption module

Based on the definition of modal fitness, we give a way of learning modal fitness through evaluating the KL divergence between unimodal distributions in text and image modal variational auto-encoders. Due to the distinct features obtained from different unimodal samples, we utilize a variational

encoder to capture the distribution of unimodal features. We sample the unimodal features, including e_s^T and e_s^I, in a potential space with an isotropic Gaussian prior. The variational posterior probability of text or image modal observation can be expressed as follows:

$$q\left(z|j\right) = \mathcal{N}\left(z|\mu\left(j\right), \sigma\left(j\right)\right), \tag{5}$$

where the mean μ and the variance σ can be obtained from modality-specific encoder of text or image modal. More generally, given each data sample x_s which contains aligned text feature vector e_s^T and image feature vector e_s^I. And the variational posterior for both modalities is defined as follows:

$$\begin{cases} q\left(z_e^T \mid e_s^T\right) = \mathcal{N}\left(z_e^T \mid \mu\left(e_s^T\right), \sigma\left(e_s^T\right)\right) \\ q\left(z_e^I \mid e_s^I\right) = \mathcal{N}\left(z_e^I \mid \mu\left(e_s^I\right), \sigma\left(e_s^I\right)\right) \end{cases}. \tag{6}$$

We obtain the final distribution based on all samples as follows:

$$\begin{cases} q\left(z^T\right) = \dfrac{1}{N}\sum_{s=1}^{N} q\left(z_s^T \mid e_s^T\right) \\ q\left(z^I\right) = \dfrac{1}{N}\sum_{s=1}^{N} q\left(z_s^I \mid e_s^I\right) \end{cases}. \tag{7}$$

The alignment of distributions for two modalities can be quantified by the following two KL divergences as follows:

$$\begin{cases} \varepsilon_s^{TI} = \dfrac{D_{KL}\left(q\left(z_e^T \| e_s^T\right) \| q\left(z_e^I \| e_s^I\right)\right)}{D_{KL}\left(q\left(z^T\right) \| q\left(z^I\right)\right)} \\ \varepsilon_s^{IT} = \dfrac{D_{KL}\left(q\left(z_e^I \| e_s^I\right) \| q\left(z_e^T \| e_s^T\right)\right)}{D_{KL}\left(q\left(z^I\right) \| q\left(z^T\right)\right)} \end{cases}, \tag{8}$$

where D_{KL} denotes the KL divergence between text and image, and the fitness score ε_s is the symmetrized KL divergence based on averaging over the normalized values of $D_{KL}\left(q\left(z_s^T \| e_s^T\right) \| q\left(z_s^I \| e_s^I\right)\right)$ and $D_{KL}(q\left(z_s^I \| e_s^I\right) \| q\left(z_s^T \| e_s^T\right))$.

We use the modal fitness score ε_t and ε_t as the weight to adaptively adjust the relationship between unimodal features and cross-modal fusion features. The final modal fitness score for the two modalities is obtained by these KL divergence values as follows.

$$\begin{cases} \varepsilon_t = sigmoid\left(\varepsilon_s^{TI}\right) \\ \varepsilon_i = sigmoid\left(\varepsilon_s^{IT}\right) \end{cases}, \tag{9}$$

where $sigmoid$ is the activation function that aims to map the fitness score between 0 and 1. The final representation $\widetilde{x}$ consists of unimodal representations and multimodal features.

$$\widetilde{x} = \left(\varepsilon_t \times e_s^T\right) \oplus \left(\varepsilon_h \times h^{'}\right) \oplus \left(\varepsilon_i \times e_s^I\right), \tag{10}$$

where $\oplus$ represents the concatenation operation and $\varepsilon_t + \varepsilon_h + \varepsilon_i = 1$. Then we put $\widetilde{x}$ into the fully-connected networks to prediction label.

$$\widetilde{y} = softmax\left(MLP\left(\widetilde{x}\right)\right). \tag{11}$$

Considering that fake news detection is a typical binary classification task, we introduce the cross-entropy function as a loss function for classification.

$$\mathcal{L}_2 = ylog\left(\widetilde{y}\right) + \left(1 - \widetilde{y}\right) log\left(1 - y\right), \tag{12}$$

where y represents the ground-truth label. We control the effect of semantic alignment on the classification task by setting $\gamma \in (0, 1)$ on the loss function, and the final loss function of ILMF-FND can be expressed as:

$$\mathcal{L} = \gamma\mathcal{L}_1 + \mathcal{L}_2. \tag{13}$$

Algorithm 1 The procedure of ILMF-FND

Require: The original dataset $\mathcal{D}$. The model set $\mathcal{X} = \left\{x^i{}_1, x^i{}_2, \cdots, x^i{}_N\right\}$ and the ground-truth label set $\mathcal{Y} = \{y_1, y_2, \cdots, y_N\}$.

Ensure: The predicted news label $\widetilde{y}_n \in \widetilde{\mathcal{Y}}$ of dataset $\mathcal{D}$ on both Twitter and Weibo.

for $\{x, y\} = \left\{\left\{x^i\right\}_n, y\right\}$ in $\mathcal{D}$ **do**
  Unimodal embedding: $e^T \longleftarrow x^T$.
  Get aligned vectorse$_s^T$ and e_s^I.
  $Z_e^T = \sigma(W_4[h_{t-1}, e_s^T] + b_4) * tanh(\sigma(\left(W_1 \cdot \left[h_{t-1}, e_s^T\right] + b_1\right)) * C_{t-1} + \sigma\left(W_2 \cdot \left[h_{t-1}, e_s^T\right] + b_2\right) * tanh\left(W_3 \cdot \left[h_{t-1}, e_s^T\right] + b_3\right))$. ▷ Eq. (2)
  $Z = \begin{bmatrix} Z_e^T \\ 1 \end{bmatrix} \otimes \begin{bmatrix} Z_e^I \\ 1 \end{bmatrix}$. ▷ Eq. (3)
  $h = \left(\sum_{i=1}^{r} \otimes_{j=1}^{2} W_j^{(i)}\right) \cdot Z$. ▷ Eq. (4)
  Generate assigned weights: $\varepsilon_t, \varepsilon_i, \varepsilon_h \longleftarrow$ KL divergence.
  $\widetilde{x}_n = \left(\varepsilon_t \times e_s^T\right) \oplus \left(\varepsilon_h \times h\right) \oplus \left(\varepsilon_i \times e_s^I\right)$. ▷ Eq. (10)
  for n in $[1, N]$ **do**
    $\widetilde{y}_n = softmax\left(MLP\left(\widetilde{x}_n\right)\right)$. ▷ Eq. (11)
  end for
  $\widetilde{\mathcal{Y}} = \{\widetilde{y_1}, \widetilde{y_2}, \cdots, \widetilde{y_N}\}$.
  return $\widetilde{\mathcal{Y}}$.
end for

ILMF-FND algorithm

We clearly list the important steps of ILMF-FND in Algorithm 1. Given a dataset $\mathcal{D}$: $\{x, y\} = \left\{\left\{x^i\right\}_n, y\right\}_N$, where N

Table 1 Statistics of the datasets used in our experiments

Statistics	Twitter	Weibo
# fake news on the training set	5007	3749
# real news on the training set	6840	3783
# news on the test set	1406	1996

is sample size. We first obtain the unimodal embedding e^T and e^I via BERT and ResNet-34 in top two step. Then we get the aligned vectors e_s^T and e_s^I in step 3. Next, both vectors are fed into *Cross-modal Fusion Module* to obtain the final compressed representation h, as shown in steps 4–6. Note that, the steps above take the text modal as an example, and similar steps are performed on the image modal. After that, we quantify modal fitness via KL divergence and adaptively obtain final assigned weights in step 7. Therefore, the n-th representation $\widetilde{x}_n$ is composed of unimodal and multimodal embedding. Specifically, we multiply the unimodal and multimodal embedding with their weights respectively, and then perform a concatenation operation. Then, we input the n-th representation $\widetilde{x}_n$ into the fully-connected networks to predict the n-th label $\widetilde{y}_n$. Finally, we repeat these steps above to get all of their labels and return the prediction label set $\widetilde{\mathcal{Y}}$.

Experiments

In this section, we present the experimental configurations in "Experimental configurations" section first. Then, we discuss the overall performance in "Results and analysis" section. Next, the ablation studies are done in "Ablation studies" section, including the effectiveness of each component and quantitative analysis. Besides, we perform the convergence analysis to verify the overfitting risk of our model in "Convergence analysis" section and conduct model parameter trials in "Model parameter trials" section. Finally, we organize efficiency analysis experiments and case studies in "Efficiency analysis" and "Case studies" sections.

Experimental configurations

Datasets

There are two widely used datasets from social media, as shown in Table 1.

1) ***The Twitter Dataset***.[1] Twitter was released for MediaEval Verifying Multimedia Use task [39]. Due to the fact that only text and image are considered in this paper, we filter posts which include videos based on existing works. In our experiments, we maintain the same data split scheme according to the benchmark [39], i.e., there are 6840 real news and 5007 fake news in training set, and the test set totally contains 1406 posts.

2) ***The Weibo Dataset***.[2] Weibo was published by [28]. In our experiments, the same data split scheme [28] is maintained as in previous work, i.e., there are 3749 fake news and 3783 real news in training set, and the test set totally contains 1996 posts.

Both datasets we discuss above have been data preprocessed so that each text has at least a corresponding image. Besides, as previous researches do [24, 40], we delete those news which have just text or images because our model is focused on multimodal news.

Baselines

The baselines listed here include both unimodal and multimodal methods. where the top two are unimodal methods, and the ones after these are both multimodal methods.

- **CAR** [15]: A plain text detection method that synthesizes RNN and attention mechanisms.
- **VS** [18]: A detection method combining image content and statistical features.
- **RA** [28]: A detection approach for fake news consists of LSTM network and attention mechanism.
- **EANN** [41]: EANN utilizes a multimodal feature extractor and a fake posts detector to support fake news detection.
- **MVAE** [42]: MVAE uses variational auto-encoders to model images and text and achieve classification.
- **MKEMN** [27]: MKEMN fuses text, images and retrieved knowledge embeddings through convolutional operations.
- **MCAN** [40]: MCAN stacks multiple layers of common concern to fuse multimodal features.
- **CAFE** [24]: CAFE can adaptively aggregate unimodal and multimodal features with the help of cross-modal ambiguity.
- **CSFND** [31]: CSFND introduces contextual information into the representation learning to aid in fake news detection.
- **COOLANT** [26]: COOLANT first adds auxiliary tasks for the purpose of softening the negative sample loss in contrast learning.

[1] https://www.dropbox.com/s/7ewzdrbelpmrnxu/rumdetect2017.zip?dl=0.

[2] https://drive.google.com/file/d/14VQ7EWPiFeGzxp3XC2DeEHi-BEisDINn/view?usp=sharing.

Research questions

In order to further analyze the performance of our proposed model ILMF-FND, we pay attention to each of the following research question.

- **RQ1** Does our proposed model ILMF-FND significantly improve the performance for the task of fake news detection compared to baselines on both Twitter and Weibo datasets?
- **RQ2** which module of our model plays an vital role in overall performance?
- **RQ3** Can the contribution of each module to the ILMF-FND be made more visible through visualization?
- **RQ4** Does ILMF-FND face the risk of overfitting during operation?
- **RQ5** How is the relationship between the number of parameters and the performance of ILMF-FND compared to baselines?
- **RQ6** Does our model outperform other baselines in terms of memory size and the runtime?
- **RQ7** Can some cases be given to show how ILMF-FND specifically identifies fake news?

Implementation details

We specify that the text encoder input should not exceed 200 words in length. We then encode the text by a pre-trained BERT [33] model with a dimension of 256, i.e, $d_1 = 256$. Similarly, we convert 224×224 images into feature vectors by pre-trained ResNet-34 [34] and the dimension of feature vectors is 512, i.e, $d_2 = 256$. The obtained feature vectors are fed into three fully connected layers which have 64 hidden units in each layer to realize the MLP for each modality. In the cross-modal fusion module, the ILMF [14] mechanism is introduced. And the modal-aligned feature vectors are first fed into the LSTM [36] to compress the time series information and extract the hidden states, whereas the low-rank tensor is used to realize the cross-modal fusion. The weights are redistributed according to the modal fitness to complete the fake news classification finally.

In all experiments in this paper, the parameter margin m is 0.2 in Eq. (1), the hyperparameter γ is set to 0.6 in Eq. (13), and the initial learning rate is set to 10^{-4}. Besides, the same data split scheme is maintained for all method comparisons, following the principle of the control variable method. Only the corresponding module is deleted and retrained to obtain the comparison results when performing ablation experiments. The ReLU mentioned in the paper defaults to the activation function if not specified and the epoch is always 100. In addition, we generally use Accuracy (**Acc**), Precision (**P**), Recall (**R**) and F_1 as evaluation indicators.

Results and analysis

In Tables 2 and 3, we compare ILMF-FND (ours) with other methods on Twitter and Weibo. The outcomes indicate that ILMF-FND outperforms almost all other methods in terms of Acc and F_1. Specifically, ILMF-FND achieves the highest accuracy of 90.2 and 93.4% on the two widely used datasets respectively. Besides, its Precision, Recall, and F_1-score also exceed most of the compared methods. However, one flaw we can see on the Twitter dataset is that the P-indicator is significantly lower than that of the optimal baseline.

This is mainly due to that our proposed method mainly utilizes the contrastive learning and adaptive processing of the relations between unimodal and multimodal features for fake news detection, which makes us to pay more attention to "Rumor" relative to "Non Rumor".

Since most news in twitter is related to some specific events, it leads to difficulty to identify the real news.

In this case, we focus our breakthroughs on the detection of fake news, that is why there is a gap between the p-indicator in "Rumor" and "Non Rumor" and an increase in the overall accuracy.

Both Tables 2 and 3 shows that unimodal detection methods achieve good accuracy additionally. Multimodal methods fuse modalities, compress time series features and capture hidden information to achieve even better accuracy. However, it is important to note that unimodal methods still play a significant role in multimodal conditions.

In addition, since many posts on Twitter are only related to a specific event and there is not much correlation between them, it is easy for overfitting to occur. In contrast, our model performs better on both datasets. Aligning the modal feature vectors to closely link the two modalities can significantly improve the recognition of a single event and enhance detection performance. The subsequent cross-modal fusion further improves the fusion fitness of the feature vectors, while simultaneously reducing parameters. This approach combines the modal fitness quantitative measure of the difference between unimodal with adaptive assignment of weights, resulting in a more efficient model.

Ablation studies

To validate the components of the ILMF-FND model, we develop two distinct groups of experiments.

Effectiveness of each component

We aim to evaluate the influence of each component of ILMF-FND on fake news detection in the initial experiment. The control variable method is used to sequentially input the variants formed by deleting a certain part of the proposed

Table 2 Performance comparison between ILMF-FND and other methods on Twitter dataset

Method	Acc	Rumor			Non rumor		
		P	R	F_1	P	R	F_1
CAR	0.637	0.574	0.690	0.682	0.724	0.602	0.617
VS	0.617	0.635	0.644	0.639	0.639	0.630	0.634
RA	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
MKEMN	0.715	0.814	0.756	0.708	0.634	0.774	0.660
MCAN	0.809	0.889	0.765	0.822	0.732	0.871	0.795
CAFE	0.806	0.807	0.799	0.803	0.805	0.813	0.809
CSFND	0.833	0.899	0.799	0.846	0.763	0.878	0.817
COOLANT	0.900	0.879	**0.922**	**0.900**	**0.923**	0.880	0.901
ILMF-FND	**0.902**	**0.950**	0.852	0.898	0.861	**0.954**	**0.905**

We underline the optimal baseline and bold the optimal result

Table 3 Performance comparison between ILMF-FND and other methods on Weibo dataset

Method	Acc	Rumor			Non rumor		
		P	R	F_1	P	R	F_1
CAR	0.745	0.705	0.765	0.750	0.756	0.725	0.740
VS	0.726	0.732	0.712	0.722	0.720	0.740	0.730
RA	0.772	0.854	0.656	0.742	0.720	0.889	0.795
EANN	0.795	0.806	0.795	0.800	0.752	0.793	0.804
MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
MKEMN	0.814	0.823	0.799	0.812	0.723	0.819	0.798
MCAN	0.899	0.913	0.889	0.901	0.884	0.909	0.897
CAFE	0.840	0.855	0.830	0.842	0.825	0.851	0.837
CSFND	0.895	0.899	0.895	0.897	0.892	0.896	0.894
COOLANT	0.923	0.927	0.923	0.925	0.919	0.922	0.920
ILMF-FND	**0.934**	**0.944**	**0.934**	**0.929**	**0.942**	**0.924**	**0.933**

We underline the optimal baseline and bold the optimal result

ILMF-FND model into the same training and test sets, and compare the final results.

Specifically, the various types of variants are listed below: *1) ILMF-FND w/o IFA*, which removes the modal information feature alignment module and directly utilizes the unimodal feature vectors as subsequent inputs. *2) ILMF-FND w/o CMF*, which removes the cross-modal fusion module and uses a simple sum of the text and image feature vectors as a substitute. And *3) ILMF-FND w/o MAM*, which removes the modal adaption module, does not use an adaptive adjustment of the weighting methods that simply assumes that multimodal features are equally important as each unimodal feature.

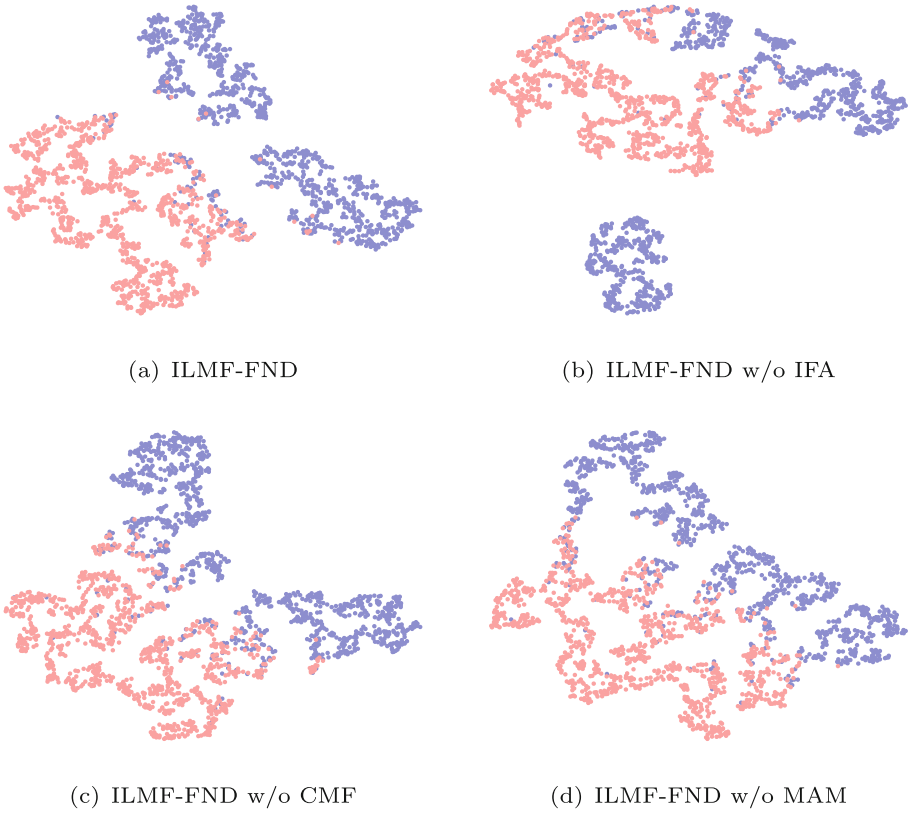
Table 4 illustrates the results of the ablation studies. Overall, the variants with a particular module removed all performed lower than the original ILMF-FND, which illustrates the validity of each component in the model. In particular, we find and summarise the following points:

Table 4 Ablation studys and efficacy analysis are conducted on modules in ILMF-FND on two datasets

	Method	Acc	F_1	
			Rumor	Non rumor
Twitter	ILMF-FND	**0.902**	**0.898**	**0.905**
	-w/o IFA	0.885	0.874	0.894
	-w/o CMF	0.890	0.882	0.897
	-w/o MAM	0.888	0.876	0.885
Weibo	ILMF-FND	**0.934**	**0.929**	**0.933**
	-w/o IFA	0.921	0.920	0.922
	-w/o CMF	0.928	0.927	0.930
	-w/o MAM	0.928	0.927	0.929

We bold the optimal result

- The poor performance of ILMF-FND w/o IFA compared to the other variants on both datasets suggests that it is particularly important to use the eigenvectors obtained



(a) ILMF-FND (b) ILMF-FND w/o IFA

(c) ILMF-FND w/o CMF (d) ILMF-FND w/o MAM

Fig. 4 T-SNE visualization of features learned by ILMF-FND and three variants on the Weibo test dataset. Dots of the same color belong to the same label

by aligning the information features of each modality as inputs to subsequent modules. Information alignment can map the feature vectors of each modality into a shared semantic space, refining the features and reducing the interference of noise on later cross-modal fusion.

- ILMF-FND w/o CMF and ILMF-FND w/o MAM have a more significant impact on performance in the Twitter dataset compared to the Weibo dataset. The data processing produces different noises due to the fact that they come from different social platforms, which interferes with the performance of model. Additionally, Twitter has a word limit for texts, resulting in a fragmented dataset that mostly focuses on a specific event and topic. Our experiments confirm the importance of cross-modal fusion and modal adaption in processing posts related to a single event.

Quantitative analysis

To make the experimental results clearer and more observable, we analyze them further after visualizing them with T-SNE [43]. Taking the Weibo dataset as an example, ILMF-FND can extract these features and its variants with visual representation, as shown in Fig. 4.

Figure 4 suggests that the boundaries of the different coloured points are more distinct in Fig. 4a compared to the remaining three variations, which means that our model can extract sharper features with greater accuracy. Figure 4b lacks the information alignment module and does not map the feature vectors of text and image into a same semantic space, which results in the inability to effectively learn multimodal features in depth. In detail, the sense of secession of dots with the same label is more obvious so that the effect of the present model is more reduced in comparison to the original model. The borders of the red and blue dots in Fig. 4c, d seem to be ambiguous, which also show that the efficient and accurate multimodal method can obviously enhance the representation of the fusion features. Thus, we find that the adaptive weight assignment scheme further improves the model performance. Overall, through the intuitive visualization results we can conclude that each part of the model is contributory and irreplaceable.

Convergence analysis

Convergence is crucial for a model, which determines whether it will suffer from gradient vanishing, gradient exploding or overfitting. To explore and analyze whether the ILMF-FND model faces overfitting during training, we generate average loss curves of the ILMF-FND across Twitter and Weibo datasets, as shown in Fig. 5.

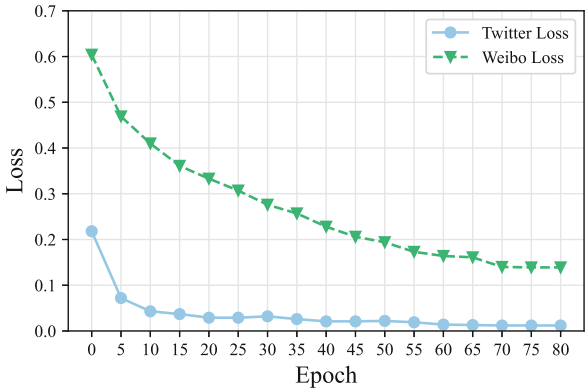


Fig. 5 Average loss curves for Twitter and Weibo datasets. Only the loss values of the top 80 epochs are shown here because the difference in the loss value of the subsequent epochs is negligible

It is suggested that both the Twitter and Weibo loss values decline at the outset progressively and gradually stabilize later in the experiments. Specifically, our model actually stabilizes after the top 80 epochs although we set the epoch as 100, which means that our model reaches a certain equilibrium. We introduce a contrastive learning task and adopt an early stop policy to prevent the problem of overfitting. In detail, for reducing the impact on the test object, the contrastive learning task is capable of extracting event-invariant features. Therefore, our model makes a certain improvement in the effectiveness of fake news detection for new posts.

Additionally, we minimize the number of parameters to help control the complexity of our model via the modal adaption module. Finally, we divide the loss function into two parts and design an adaptive mechanism. This mechanism not only optimizes our model but also manages the issue of overfitting, which allows our model to converge relatively rapidly.

Model parameter trials

A model takes less time to run and requires less memory for the same amount of time when there are fewer parameters. However, we must ensure the quality of the model before reducing the number of parameters. To verify the impact of parameter size on model performance in ILMF-FND, we still use Twitter and Weibo to compare ILMF-FND with three other recent baselines. Figure 6 exhibits our experimental results. We reflect the relationship between model performance and the number of parameters more clearly by keeping the evaluation indicator values and the number of parameters on the same graph.

We find that ILMF-FND uses fewer parameters to achieve optimal performance on both Twitter and Weibo in Fig. 6. COOLANT performs similarly to ILMF-FND on the Twitter dataset while its number of parameters increases geometrically on the Weibo dataset to achieve better performance. Its performance seems instable. Similarly, CAFE has fewer parameters but poorer performance. CSFND has average performance but a large number of parameters. It can be seen that ILMF-FND is able to improve detection results while significantly reducing the number of parameters, which effectively saves costs.

Efficiency analysis

Simultaneously, we also compare ILMF-FND with the baselines in terms of the memory size and the runtime, as shown in Table 5. It can be seen that both the time cost and the memory cost, our model outperforms the baselines on both Twitter and Weibo. Only from the data, CAFE seems to have less time and memory, but its performance is somewhat lower than ours. By comparing the results with those of "Model parameter trials" section, we can see that the results of the efficiency analysis are generally consistent with the results of the model parameter trials. This is due to the introduction of ILMF in cross-modal fusion, which can effectively reduce the number of parameters and improve the fusion efficiency. Specifically, the decrease in the number of parameters directly leads to a reduction in the memory and time of our model. Overall, our proposed model ILMF-FND performs better than other models in three aspects: the performance, the memory size and the runtime.

Case studies

We compare our model with others in the test set of Twitter and Weibo to validate the capability of the approach employed in fake news detection. As shown in the Fig. 7, there are three fake cases that can be captured by ILMF-FND but are not fully and accurately recognized by existing methods.

To show the detection performance of our model clearly, we list the recognition results of ILMF-FND and baselines in the Table 6. These cases are correctly identified mainly due to the adaptive weight assignment and the processing of unimodal and cross-modal feature relationships in our model. For Fig. 7a, we can hardly determine whether it is a fake case just from the textual contents. Similarly, the image in Fig. 7b shows no signs of forgery while the text seems unbelievable. However, our model can find the attached image in Fig. 7a is a forged picture because of the higher weight assigned to the image modal and also can identify the textual contents in Fig. 7b as a fake case via the higher weight of the text modal. In contrast, if our model considers only multimodal fusion features and ignores unimodal features, then the both two cases are likely to be considered true. We consider that such advantages are due to the adaptive weight assignment mechanism,

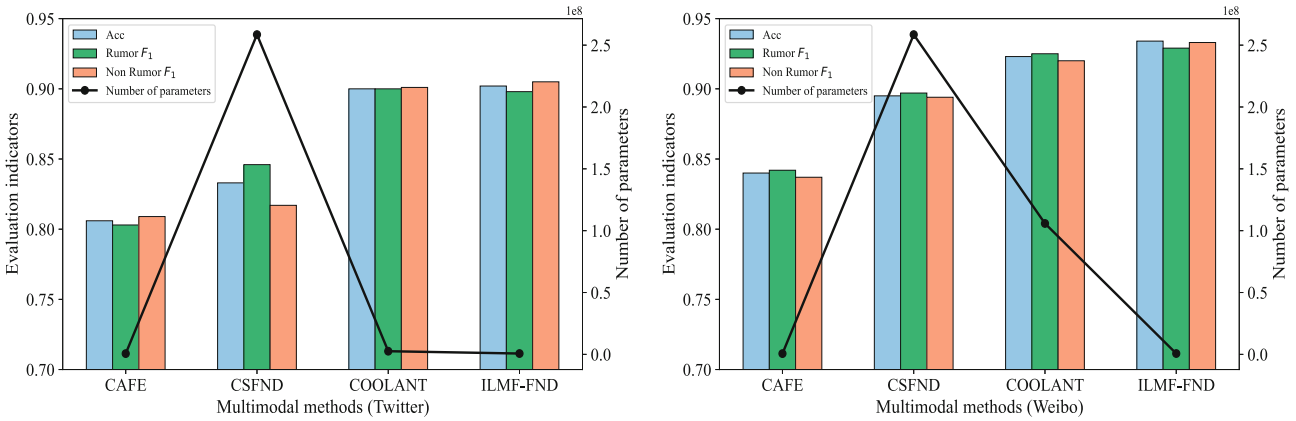


Fig. 6 Results of multimodal detection methods on Twitter and Weibo datasets, where bar graphs indicate accuracy, rumor F_1 and non rumor F_1 respectively; line graphs indicate the number of parameters

Table 5 Efficiency comparison between ILMF-FND and the baselines

Dataset	Metric	CAFE	CSFND	COOLANT	ILMF-FND
Twitter	Time (*s*/epoch)	9.70	13.25	23.45	13.00
	Memory (GB)	1.74	6.67	7.04	1.59
Weibo	Time (*s*/epoch)	16.14	18.04	17.19	15.15
	Memory (GB)	1.49	5.60	1.70	1.53

(a) Lenticular clouds over Mount Fuji. (b) A woman gives birth to 14 children from 14 different fathers. (c) Syrian girl selling chewing gum in the street of Jordan.

Fig. 7 Three fake cases correctly detected by ILMF-FND. The fake types can be summarised from left to right as Tampered Image, Fake Text and False Connection

which can optimize the model by dynamically adjusting the weights of the multimodal and unimodal features. Unlike our model, the CAFE and the COOLANT focus on cross-modal features so that they are difficult to detect anomalies in unimodal.

In addition, it is hard to judge Fig. 7c whether it is a fake case just from its text and image. But our model assumes that it is fake because there is no authenticity association built between text and image. Therefore, the adaptive weight assignment mechanism assigns higher weights to cross-modal features than to unimodal features. The judgement made by our model is reasonable. We can notice that the image describes a happy emotion that has nothing to do with the war atmosphere rendered by the text, which is a false case due to false connection. For the CSFND, we believe that its inability to recognize this case is due to an excessive focus on unimodal content at the expense of inter-modal connections. To summarize, our model ILMF-FND is effective in identifying different categories of multimodal fake news.

Table 6 A comparison of the recognition results between ILMF-FND and other baselines for these fake cases in Fig. 7, where ✓ stands for the method can recognize the case and ✗ means that it cannot

Methods	Figure 7a	Figure 7b	Figure 7c
CAFE	✗	✗	✓
CSFND	✓	✓	✗
COOLANT	✗	✗	✓
ILMF-FND(ours)	✓	✓	✓

Conclusion

In this paper, we propose a novel cross-modal fusion method ILMF-FND based on information enhancement for multimodal fake news detection. Unlike previous methods, we do information enhancement for text and images before semantic alignment to extract hidden information and complete denoising. We also achieve the best performance by improving the LMF while minimizing the number of parameters. Finally, an adaptive mechanism is designed to optimize the model by dynamically adjusting the weights of the multimodal and unimodal features according to the current modal fitness. ILMF-FND demonstrates superior performance over existing models in terms of both accuracy and parameter size, as evidenced by results from the Twitter and Weibo datasets. Besides, the source code of our paper can be found in https://github.com/asufdahu/ILMF-FND.

For subsequent endeavors, to begin with, we intend to add a knowledge graph into our model, as the knowledge graph [44] that includes more real-world knowledge can make test results more in line with reality. In addition, we would like to optimize the acquisition of unimodal embedding from news to further improve the quality of fake news detection. Alternatively, we aim to examine our proposal on other datasets where the labels are available.

Acknowledgements The anonymous reviewers are acknowledged for their constructive comments. The authors would like to thank the support by the COSTA: complex system optimization team of the College of System Engineering at NUDT.

Data availability Data will be made available on request.

References

1. Conroy NK, Rubin VL, Chen Y (2015) Automatic deception detection: methods for finding fake news. Proc Assoc Inf Sci Technol 52(1):1–4
2. Campbell M (2016) Meeting the challenges of the fake news epidemic. Arab American News, 9
3. Liu H, Chen S, Cao S, Zhu J, Ren Q (2023) A study on fake news detection based on multimodal learning. J Front Comput Sci Technol 2015–2029
4. Masciari E, Moscato V, Picariello A, Sperlì G (2020) A deep learning approach to fake news detection. In: ISMIS. Lecture Notes in Computer Science, vol. 12117, pp. 113–122
5. Ghamdi MAA, Bhatti MS, Saeed A, Gillani Z, Almotiri SH (2024) A fusion of Bert, machine learning and manual approach for fake news detection. Multimed Tools Appl 83(10):30095–30112
6. Zhou X, Zafarani R (2019) Network-based fake news detection: a pattern-driven approach. ACM SIGKDD Explor Newslett 21(2):48–60
7. Liu X, Nourbakhsh A, Li Q, Fang R, Shah S (2015) Real-time rumor debunking on twitter. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp. 1867–1870
8. Pérez-Rosas V, Kleinberg B, Lefevre A, Mihalcea R (2017) Automatic detection of fake news. CoRR arxiv:1708.07104
9. Mridha MF, Keya AJ, Hamid MA, Monowar MM, Rahman MS (2021) A comprehensive review on fake news detection with deep learning. IEEE Access 9:156151–156170
10. Kumar S, Kumar S, Yadav P, Bagri M (2021) A survey on analysis of fake news detection techniques. In: 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), pp. 894–899. IEEE
11. Yu F, Liu Q, Wu S, Wang L, Tan T (2017) A convolutional approach for misinformation identification. In: IJCAI, pp. 3901–3907
12. Ying Q, Hu X, Zhou Y, Qian Z, Zeng D, Ge S (2023) Bootstrapping multi-view representations for fake news detection. In: AAAI, pp. 5384–5392
13. Baltrušaitis T, Ahuja C, Morency L-P (2018) Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell 41(2):423–443
14. Sahay S, Okur E, Kumar SH, Nachman L (2020) Low rank fusion based transformers for multimodal sequences. arXiv preprint arXiv:2007.02038
15. Chen T, Li X, Yin H, Zhang J (2018) Call attention to rumors: deep attention based recurrent neural networks for early rumor detection. PAKDD (Workshops). In: Trends and applications in knowledge discovery and data mining. PAKDD 2018. Lecture Notes in Computer Science vol. 11154, pp. 40–52. Sprimger
16. Bian T, Xiao X, Xu T, Zhao P, Huang W, Rong Y, Huang J (2020) Rumor detection on social media with bi-directional graph convolutional networks. In: AAAI, pp. 549–556
17. Li Y, Feng D, Lu M, Li D (2019) A distributed topic model for large-scale streaming text. In: Knowledge Science, Engineering and Management: 12th International Conference, KSEM 2019, Athens, Greece, August 28–30, 2019, Proceedings, Part II 12, pp. 37–48 . Springer
18. Jin Z, Cao J, Zhang Y, Zhou J, Tian Q (2017) Novel visual and statistical image features for microblogs news verification. IEEE Trans Multim 19(3):598–608
19. Zhao Y, Zobel J (2007) Searching with style: Authorship attribution in classic literature. In: Proceedings of the Thirtieth Australasian Conference on Computer science-Volume 62, pp. 59–68. Citeseer
20. Ma J, Gao W, Wong K-F (2019) Detect rumors on twitter by promoting information campaigns with generative adversarial learning. In: The World Wide Web Conference, pp. 3049–3055
21. Sun W, Chen T (2020) Research on microblog rumor recognition method based on albert-bilstm model. Comput Era 8:21–26
22. Mayank M, Sharma S, Sharma R (2022) Deap-faked: Knowledge graph based approach for fake news detection. In: 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 47–51. IEEE

23. Zheng J, Zhang X, Guo S, Wang Q, Zang W, Zhang Y (2022) MFAN: multi-modal feature-enhanced attention networks for rumor detection. In: IJCAI, pp. 2413–2419
24. Chen Y, Li D, Zhang P, Sui J, Lv Q, Lu T, Shang L (2022) Cross-modal ambiguity learning for multimodal fake news detection. In: WWW, pp. 2897–2905
25. Qi P, Bu Y, Cao J, Ji W, Shui R, Xiao J, Wang D, Chua T-S (2023) Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms. In: Proceedings of the AAAI Conference on Artificial Intelligence 37, pp. 14444–14452
26. Wang L, Zhang C, Xu H, Xu Y, Xu X, Wang S (2023) Cross-modal contrastive learning for multimodal fake news detection. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 5696–5704
27. Zhang H, Fang Q, Qian S, Xu C (2019) Multi-modal knowledge-aware event memory network for social media rumor detection. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 1942–1951
28. Jin Z, Cao J, Guo H, Zhang Y, Luo J (2017) Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: Proceedings of the 25th ACM International Conference on Multimedia, pp. 795–816
29. Castelo S, Almeida TG, Elghafari A, Santos ASR, Pham K, Nakamura EF, Freire J (2019) A topic-agnostic approach for identifying fake news pages. In: WWW (Companion Volume), pp. 975–980
30. Min E, Rong Y, Bian Y, Xu T, Zhao P, Huang J, Ananiadou S (2022) Divide-and-conquer: Post-user interaction network for fake news detection on social media. In: WWW, pp. 1148–1158
31. Peng L, Jian S, Kan Z, Qiao L, Li D (2024) Not all fake news is semantically similar: contextual semantic representation learning for multimodal fake news detection. Inf Process Manag 61(1):103564
32. Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828
33. Devlin J, Chang M, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. CoRR arxiv:1810.04805
34. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778
35. Ma J, Zhao Z, Yi X, Chen J, Hong L, Chi EH (2018) Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1930–1939
36. Baccouche M, Mamalet F, Wolf C, Garcia C, Baskurt A (2010) Action classification in soccer videos with long short-term memory recurrent neural networks. In: ICANN (2). Lecture Notes in Computer Science, vol. 6353, pp. 154–159
37. Liu Z, Shen Y, Lakshminarasimhan VB, Liang PP, Zadeh A, Morency L-P (2018) Efficient low-rank multimodal fusion with modality-specific factors. arXiv preprint arXiv:1806.00064
38. Zadeh A, Chen M, Poria S, Cambria E, Morency L-P (2017) Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:1707.07250
39. Boididou C, Papadopoulos S, Zampoglou M, Apostolidis L, Papadopoulou O, Kompatsiaris Y (2018) Detection and visualization of misleading content on twitter. Int J Multimed Inf Retr 7(1):71–86
40. Wu Y, Zhan P, Zhang Y, Wang L, Xu Z (2021) Multimodal fusion with co-attention networks for fake news detection. In: ACL/IJCNLP (Findings). Findings of ACL, vol. ACL/IJCNLP 2021, pp. 2560–2569
41. Wang Y, Ma F, Jin Z, Yuan Y, Xun G, Jha K, Su L, Gao J (2018) 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
42. Khattar D, Goud JS, Gupta M, Varma V (2019) MVAE: multimodal variational autoencoder for fake news detection. In: WWW, pp. 2915–2921. ACM
43. Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9:2579–2605
44. Zhang L, Zhang X, Zhou Z, Huang F, Li C (2024) Reinforced adaptive knowledge learning for multimodal fake news detection. In: AAAI, pp. 16777–16785