External Reliable Information-enhanced Multimodal Contrastive Learning for Fake News Detection

Biwei Cao[1, 2], **Qihang Wu**[1, 2], **Jiuxin Cao**[1, 2, 3*], **Bo Liu**[2, 3, 4], **Jie Gui**[1, 2, 3, 5*]

[1] School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China
[2] Key Laboratory of Computer Network and Information of Ministry of Education of China, Nanjing 211189, China
[3] Purple Mountain Laboratories, Nanjing 210000, China
[4] School of Computer Science and Engineering, Southeast University, Nanjing 211189, China
[5] Engineering Research Center of Blockchain Application, Supervision And Management (Southeast University), Ministry of Education
{caobiwei, wuqihang, jx.cao, bliu, guijie}@seu.edu.cn

Abstract

With the rapid development of the Internet, the information dissemination paradigm has changed and the efficiency has been improved greatly. While this also brings the quick spread of fake news and leads to negative impacts on cyberspace. Currently, the information presentation formats have evolved gradually, with the news formats shifting from texts to multimodal contents. As a result, detecting multimodal fake news has become one of the research hotspots. However, multimodal fake news detection research field still faces two main challenges: the inability to fully and effectively utilize multimodal information for detection, and the low credibility or static nature of the introduced external information, which limits dynamic updates. To bridge the gaps, we propose ERIC-FND, an external reliable information-enhanced multimodal contrastive learning framework for fake news detection. ERIC-FND strengthens the representation of news contents by entity-enriched external information enhancement method. It also enriches the multimodal news information via multimodal semantic interaction method where the multimodal constrative learning is employed to make different modality representations learn from each other. Moreover, an adaptive fusion method is taken to integrate the news representations from different dimensions for the eventual classification. Experiments are done on two commonly used datasets in different languages, X (Twitter) and Weibo. Experiment results demonstrate that our proposed model ERIC-FND outperforms existing state-of-the-art fake news detection methods under the same settings.

Extended version and code —
https://github.com/Tarasom123/ERIC-FND

Introduction

With the rapid development and widespread use of the internet, the number of internet users has been increasing, making reading news online become an important part of daily activities. The Internet provides great convenience for disseminating information on social network platforms, leading to a large amount of fake news published on the Internet for various intentions. The widespread dissemination of fake news does not only severely destroy the credibility and social influence of mainstream media but also has the potential to cause panic in society. To mitigate the negative impact caused by fake news, automated fake news detection has gradually become a research hotspot.

Early fake news detection mainly focused on propagation structures or text-based detection(Shu et al. 2017a). With the development of social network platforms, the news format has gradually shifted from pure text to multimodal content. Correspondingly, the research on fake news detection has changed to multimodal fake news detection as well. Existing multimodal fake news detection methods mainly focus on how to integrate the news representation from multimodalities to improve detection performance, instead of doing in-depth research on the inconsistencies between images and texts in fake news. Furthermore, the most multimodal methods rarely explore external information and the external information used often lacks reliability or remains static (Hu et al. 2023).

To address these issues, we propose a method that aligns news text and image information to a shared feature space through multimodal contrastive learning. Then, it obtains enhanced multimodal representations via cross-modal semantic interaction. Additionally, to deepen the model understanding of news text, descriptions of knowledge entities from Wikipedia are introduced as external reliable information. An external information enhancement module is employed, utilizing an attention mechanism to enhance the feature representation of news content with the external reliable information. Our major contributions are summarized as follows.

- We propose ERIC-FND, an External Reliable Information-enhanced multimodal Contrastive learning method for Fake News Detection, which mainly consists of three modules, namely external information enhancement module, multimodal information interactive enhancement module and adaptive fusion-based classification module. The first two modules aim to leverage external information for a deeper understanding of news content and to enhance multimodal representation based on multimodal information. The adaptive fusion based classification module attempts to adaptively fuse the

*Corresponding authors

obtained features of different dimensions, improving the classification performance and interpretability.

- To achieve a better representation of news incorporating both images and text, we utilize multimodal contrastive learning to align the news textual features with its visual features. We then use cross-modal semantic interaction method to capture the deep interactive semantic features. What is more, to enhance the model understanding of the news contents, entity-enriched information integration method is proposed to obtain the news-related external information and use this information to enhance the feature representation of news.
- Extensive experiments are conducted on two datasets to demonstrate the performance of the proposed model. The results show that our model outperforms SOTA methods.

Related Work

In the early stage, the methods are mainly based on traditional machine learning, which focuses on feature engineering to extract static features from the post propagation processes (Feng, Banerjee, and Choi 2012; Pérez-Rosas et al. 2018; Zhou et al. 2020; Chen and Guestrin 2016; Zhou and Zafarani 2019). With the development of deep learning, fake news detection research shifts towards the methods based on deep learning, due to its strong ability of handling complex data and pattern recognition. Currently, the research directions for fake news detection based on deep learning can be generally divided into two types: social context-based fake news detection and content-based fake news detection (Shu et al. 2017a).

Propagation Structure-based Fake News Detection

Fake news detection methods based on social context mostly utilize the characteristics and patterns of information transmission on social media platforms to detect fake news (Shu et al. 2017b). Researchers believe that fake news exhibits different propagation patterns and network structures compared to real news during the dissemination process (Jin et al. 2016). Ma et al. (2017) model the posts diffusion with propagation trees and propose a Propagation Tree Kernel method which calculates the similarity between the propagation tree structures of different rumors. Liu et al. (2018) applies recurrent and convolutional networks to capture both global and local variations of user characteristics along propagation paths to detect fake news. Sun et al. (2023) propose a novel hyperedge walking strategy on a meta-hyperedge graph for gaining news propagation sub-structure representations in social network. To relieve the problem of lacking in labelled dataset, Fang et al. (2023) propose a tree VAE-based sentiment propagation module to leverage the propagation structure. Lin et al. (2023) represent social media rumors as diverse propagation threads and design a hierarchical prompt encoding mechanism to learn contextual representations independent of language. The Rumor Adaptive Graph Contrastive Learning (RAGCL) (Lin et al. 2023) method is proposed to improve rumor detection by adapting graph-based learning of rumor propagation tree structures and focusing on key substructures via adaptive view augmentation.

Content-based Fake News Detection

Unimodal Fake News Detection The content-based fake news detection starts with the unimodality. Ma et al. (2016) first apply deep learning to fake news detection. Since the good sequential data modeling capabilities of recurrent neural networks (RNNs), Ma et al. employ RNNs to learn hidden layer representations and capture the content changes of related texts over time. In 2017, Yu et al. (2017) are the first to model news articles using convolutional neural networks (CNNs) and propose the CNN-based CAMI method. This method maps tweets related to a news event into a vector space, concatenates them into a matrix, and then uses a CNN to extract textual features. Inspired by the adversarial learning from Generative Adversarial Networks (GAN), Ma et al. (2019) design a generator to produce noise, making the original conversational threads more complex, thus forcing the discriminator to learn stronger rumor indicative representations from difficult samples. For more information for fake news detection, several works exploit external knowledge. Hu et al. (2021) propose an end-to-end graph neural model called CompareNet, which compares the knowledge base (KB)-based entity representations with the news contextual entity representations to capture consistency between the news content and the KB. With Combination of the user comments and user information, Tseng et al. (2022) establish fact-based associations with entities in the news content. What is more, researchers also focus on the fake image detection for the news (Gupta et al. 2013; Jin et al. 2017c; Qi et al. 2019).

Multimodal Fake News Detection With the development of information technology, multimodal news formats have become mainstream. Jin et al. (Jin et al. 2017a) first use deep learning network for multimodal fake news detection. The att-RNN is proposed to combine the multimodal features including images, texts and social context. Khattar et al. (Khattar et al. 2019a) use a bimodal variational autoencoder coupled with a binary classifier to learn probabilistic latent variable models. Chen et al. (Chen et al. 2022a) propose an ambiguity-aware multimodal fake news detection method which estimate the multimodal ambiguity and capture the cross-modal correlations. Hu et al. (Hu et al. 2023) extract the rumor evidence of different modalities from the Internet and construct a multimodal dataset with the Internet search results. However, the extensive searching on the internet introduces a large amount of noise. Zhang et al. (Zhang et al. 2024) utilize reinforcement learning to build a knowledge subgraph for each news to keep the useful knowledge related to news.

Problem Formulation

Input: Given the news data set S, and n pieces of data, we have $S = \{(t_1, v_1, l_1), \ldots, (t_k, v_k, l_k), \ldots, (t_n, v_n, l_n)\}$, where$(t_k, v_k, l_k) \in S$ is the kth piece data in the dataset, t_k denotes the k-th piece of news text, v_k denotes the image

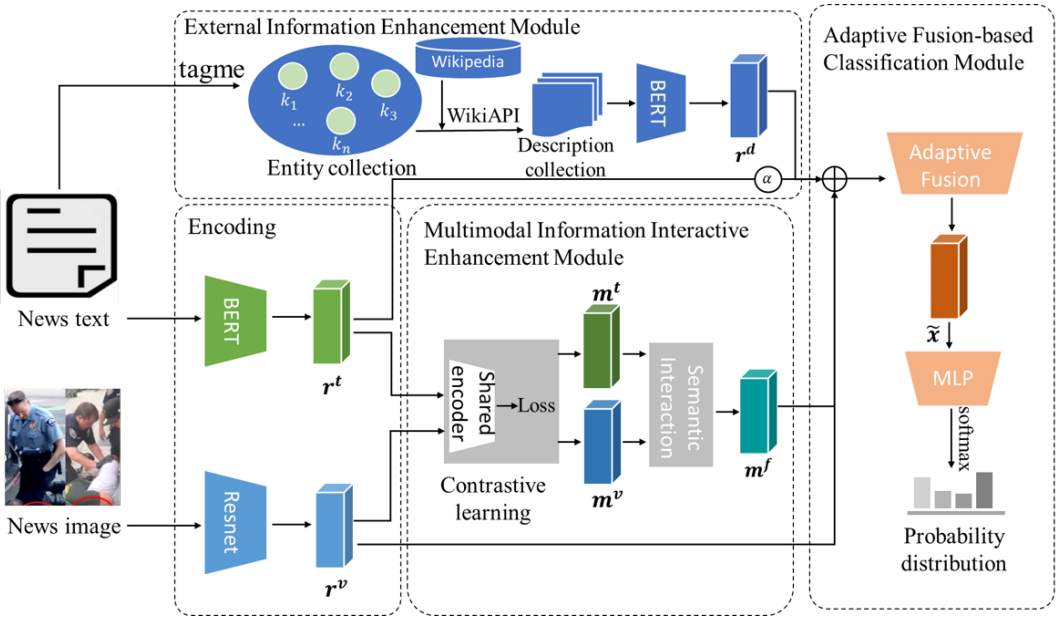


Figure 1: The proposed model ERIC-FND framework, where ⓐ represents the attention mechanism and ⊕ represents the concatenation operation.

corresponding to the k-th piece of news text, $l_k \in \{0, 1\}$ denotes the label of the k-th news item, and one piece of news corresponds to one label.

Output: A prediction label for the news, either fake news or true news.

Methodology

The framework of proposed ERIC-FND model is illustrated in Figure 1. The model has three main modules: External Information Enhancement Module to enhance the news textual content based on the external information, Multimodal Information Interactive Enhancement Module to achieve the combined multimodal representation, and Adaptive Fusion-based Classification Module to adaptively fuse all the features to complete the classification.

Encoding

We first do the feature encoding separately for the text and image of a piece of news.

For text part, we apply BERT model as the encoder on the pre-processed texts and input the vector of $[CLS]$ in the last hidden layer into the fully connected layer to obtain the representation of news text r^t, which can be formulated as follows:

$$r^{cls} = Text\_Encoder(t), \tag{1}$$

$$r^t = W_{tf} \cdot r^{cls}, \tag{2}$$

where $r^t \in \mathbb{R}^{n \times l \times d_h}$, l is the text length and d_h is the dimensionality of hidden layer. W_{tf} is the weight matrix of fully connected layer during the news text encoding process.

For image part, we start with resizing of image into the standard size. Then, the convolutional layers of ResNet-50 are used and a fully connected layer is added after the last convolutional layer. The output from the fully connected layer is taken as the representation of the news image r^v which can be formulated as follows:

$$r^{vis} = Visual\_Encoder(t), \tag{3}$$

$$r^v = W_{tf} \cdot r^{vis}, \tag{4}$$

where $r^v \in \mathbb{R}^{n \times d_v}$ and d_v is the dimensionality of image feature. W_{vf} is the weight matrix of fully connected layer during the news image encoding process.

External Information Enhancement Module

External information enhancement module consists of two parts, namely external information extraction and information enhancement which takes entity-enriched external information enhancement method to improve the news text representation by integrating with external reliable information.

External Information Extraction In the beginning, we extract the entity from news textual content to explore the entities that have the potential to be knowledge. Since X and Weibo datasets which are in different languages are selected in implementation, for Chinese, jieba segmentation tool is employed to classify the parts of speech in the news and extract noun entities from the news text, while for English, tagme tool is selected to extract all the concept entities. The knowledge concept entity collection obtained can be formulated as $E = \{k_1, k_2, k_3, \ldots, k_i, \ldots, k_n\}$, where k_i stands for the ith concept entity and n represents the number of entities.

Then, we link the news entities with Wikipedia. The extracted knowledge entities from the news text are mapped to Wikipedia and the corresponding entity descriptions are

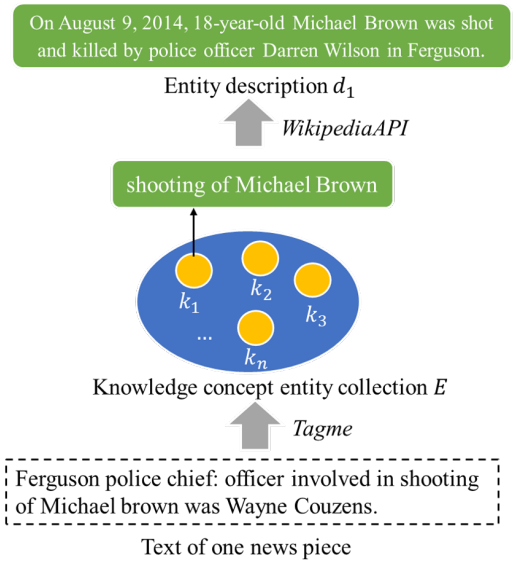


Figure 2: The process of entity description extraction.

retrieved. The specific process for obtaining entity descriptions is shown in Figure 2. The WikipediaAPI is taken to obtain the description of the knowledge concept entity "shooting of Michael Brown" from Wikipedia.

In practice, we observe that the entity descriptions returned by Wikipedia are usually long texts, where the first sentence often provides an essential conceptual description of the entity. Therefore, to avoid noise from the long entity descriptions, this study uses only the first sentence of the entity description. Additionally, when analyzing news texts, multiple entities are often identified. Thus multiple entity descriptions are retrieved from Wikipedia. Therefore, the collection D of all entity descriptions in a piece of news can be represented as $D = \{d_1, d_2, \ldots, d_i, \ldots, d_n\}$, where d_i is the description of the ith entity in Wikipedia, and n denotes the number of entity descriptions.

Attention-based Information Enhancement In this part, we do the feature extraction on the entity description collection and integrate the context feature of external information with news content.

Similar to news text encoding, we apply BERT to obtain the contextual representation of news entity description. For an entity description d, the process of getting the contextual feature r^d can be formally represented as

$$r^{desc} = Text\_Encoder(d), \tag{5}$$

$$r^d = W_{df} \cdot r^{desc}, \tag{6}$$

where W_{df} represents the weight matrix of the fully connected layer. As a result, we can achieve the entity description feature matrix of news text M^d.

Then, the representation of entity description is enhanced based on news text via attention mechanism. We calculate the dot product between news text feature r^t and the entity description feature matrix M^d as the attention weight $attF$. Next, use $attF$ to extract key features from the entity description matrix, and obtain the key entity description feature $\tilde{r}^{desc}$ by a mean pooling operation. This calculation process can be formulated as follows:

$$attF = softmax(r^t M^d), \tag{7}$$

$$\tilde{r}^{desc} = Mean\_Pooling(attF \odot M_d). \tag{8}$$

Finally, to further enhance the representation ability of news text features, we combine the obtained entity description feature $\tilde{r}^{desc}$ with the news text feature r^t through an additive fusion operation. This process results in external information enhanced text features $\tilde{r}^t$:

$$\tilde{r}^t = W_t r^t + W_d \tilde{r}^{desc}. \tag{9}$$

Multimodal Information Interactive Enhancement Module

This module has two main parts: multimodal contrastive learning and cross-modal semantic interaction. We align the multimodal news informaion through multimodal contrastive learning and gain the enhanced news representation based on the different modal features via cross-modal semantic interaction.

Multimodal Contrastive Learning Feature alignment helps balance the contributions of different modalities and reduces the loss of information. Here we apply cross-modal contrastive learning to map each unimodal feature to a shared representation space to achieve alignment of different modal features.

In the training of multimodal contrastive learning, for the batch with N examples $x = \{(x_i^v, x_i^t)\}_{i=1}^N$, there are $N \times N$ combinations of images and texts, including N positive pairs and $N^2 - N$ negative pairs, where x_i^v and x_i^t are the image and text of the ith pair. The training objective of contrastive learning is to align visual feature representations and textual feature representations by ensuring positive pairs are closer in the shared representation space compared to negative pairs.

Specifically, we first map the news text representation r^t and visual representation r^v to a shared vector space and use the L2 norm to get the text vectors e^t and visual vector e^v with the same dimensionality, which can be formulated as

$$e^t = Shared\_Encoder(r^t), \tag{10}$$

$$e^v = Shared\_Encoder(r^v), \tag{11}$$

where the shared encoders for text and image are fully connected layer structures to separately map the unimodal representations into shared vector space.

For the ith image and jth text in the batch, the similarity, which reflects the degree of semantic consistency between the image and text, is denoted as $p_{ij}^{v \to t}$. The calculation of the similarity is shown as

$$p_{ij}^{v \to t} = \frac{\exp(sim(e_i^v, e_j^t)/\tau)}{\sum_{j=1}^N \exp(sim(e_i^v, e_j^t)/\tau)}, \tag{12}$$

where $sim(\cdot)$ represents the dot product operation. τ is a hyperparameter of the temperature coefficient. The larger the value of τ is, the less the model penalizes difficult examples.

Similarily, the calculation of the similarity between the ith text and jth image in batch is:

$$p_{ij}^{t\to v} = \frac{\exp(sim(e_i^t, e_j^v)/\tau)}{\sum_{j=1}^{N}\exp(sim(e_i^t, e_j^v)/\tau)}. \tag{13}$$

where $sim(\cdot)$ represents the dot product operation. τ is a hyperparameter of the temperature coefficient.

We use the InfoNCE loss for training. The loss function $L^{v\to t}$ between images and texts can be formally represented as

$$L^{v\to t} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} y_{ij}^{v\to t}\log p_{ij}^{v\to t}. \tag{14}$$

Correspondingly, we have the loss function $L^{t\to v}$ between texts and images:

$$L^{t\to v} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} y_{ij}^{t\to v}\log p_{ij}^{t\to v}. \tag{15}$$

The overall contrastive loss function L_c is the average of $L^{v\to t}$ and $L^{t\to v}$:

$$L_c = \frac{1}{2}\left(L^{(v\to t)} + L^{(t\to v)}\right). \tag{16}$$

Cross-modal Semantic Interaction This part learns the deep semantic connections between different modalities. Given the aligned visual representation m^v and aligned text representation m^t, we first calculate the attention weights between different modality representations $f_{v\to t}$ and $f_{t\to v}$ which can be formulated as

$$f_{t\to v} = softmax\left(\frac{[m^t]\,[m^v]^T}{\sqrt{dim}}\right), \tag{17}$$

$$f_{v\to t} = softmax\left(\frac{[m^v]\,[m^t]^T}{\sqrt{dim}}\right), \tag{18}$$

where dim demotes the dimensionality.

Then, the attention weights $f_{v\to t}$ and $f_{t\to v}$ are utilized to update the aligned visual representation m^v and aligned text representation m^t, enabling unimodal features to learn the correlation from each other. The process is formulated below:

$$m_f^v = f_{t\to v}\times m^v, \tag{19}$$

$$m_f^t = f_{v\to t}\times m^t. \tag{20}$$

To further improve the interaction between visual features and textual features, the cross product is done to integrate complementary semantic representations and obtain the interaction matrix m^f:

$$m^f = m_f^v \otimes m_f^t. \tag{21}$$

Eventually, the multimodal interaction enhancement feature r^f is obtained from the interaction matrix m^f through the maxpooling operation followed by the multi-layer perceptron (MLP), which can be formulated below:

$$r^f = MLP(MaxPooling(m^f)). \tag{22}$$

Adaptive Fusion-based Classification Module

Since different features play different roles in the detection process of a specific news, it is necessary to assign specific weights to each feature for each news before performing the final fusion.

In detail, for the obtained external information enhanced text features $\tilde{r}^t$, the multimodal interaction enhancement feature r^f and the representation of the news image r^v, we first concatenate these three features horizontally and then compress the concatenated feature matrix using mean pooling. The attention weight of each feature is achieved by the mean pooling result input into the fully connected layer followed by the sigmoid activation function:

$$r = \tilde{r}^t \oplus r^v \oplus r^f, \tag{23}$$

$$\bar{r} = Mean\_Pooling(r), \tag{24}$$

$$a^t, a^v, a^f = sigmoid(FC(\bar{r}, W)), \tag{25}$$

where $\oplus$ denotes the concatenation operation and FC is the fully connected layer. At last, the final news representation $\tilde{x}$ for classification is gained from these feature weights:

$$\tilde{x} = (a^t \times \tilde{r}^t) \oplus (a^v \times r^v) \oplus (a^f \times r^f), \tag{26}$$

where $\oplus$ denotes the concatenation operation.

In the classification part, the final news representation $\tilde{x}$ is fed into a MLP and obtain the probability distribution $\hat{y}$ by the softmax function which is formulated as

$$\hat{y} = softmax(MLP(\tilde{x})). \tag{27}$$

We take the cross-entropy loss to measure the difference between the predicted probability distribution $\hat{y}$ and the true labels y. The loss function is shown in Equation 28:

$$L_d = -(y\log(\hat{y}) + (1-y)\log(1-\hat{y})). \tag{28}$$

In the end, the final loss function L is the combination of contrastive learning loss and the classification loss which is formulated as follows.

$$L = L_c + L_d. \tag{29}$$

Finally, training with L_c as the goal allows us to obtain aligned text vector m^t and aligned visual vector m^v.

Experiment

Experiment Setup

Datasets In this paper, we evaluate the proposed model ERIC-FND using the two widely used datasets Weibo and X.

Weibo dataset is constructed by Jin et al. (2017a) with the fake news in the dataset from misinformation collection by Weibo official from May 2012 to January 2016. The fake news data is from authoritative Chinese news agencies, such

	Real News	Fake News	Total
Training Set	3783	3745	7528
Test Set	999	996	1995
Total	4782	4741	9523

Table 1: Weibo dataset distribution.

	Real News	Fake News	Total
Training Set	5722	5530	11252
Test Set	640	622	1262
Total	6362	6152	12514

Table 2: X dataset distribution.

as Xinhua News Agency and CCTV News. The preprocessing of Weibo dataset follows Khattar's work (Khattar et al. 2019a). News pieces with videos or without text or images are excluded. The remaining data are split into a training set and a test set in an 8:2 ratio. For the condition that one news piece related to multiple images, only the first image is selected. As shown in Table 1, the dataset consists of 9,523 news pieces, with 7,528 in the training set and 1,925 in the test set. Each news item in the dataset has its corresponding images.

X dataset is used in competition MediaEval (Boididou et al. 2015) for automatically detecting fake news in various media formats on X. Similar to the Weibo dataset, only the data with both images and text are applied in the experiment. As shown in Table 2, the dataset has 12,514 news pieces, with 11,252 in the training set and 1,262 in the test set.

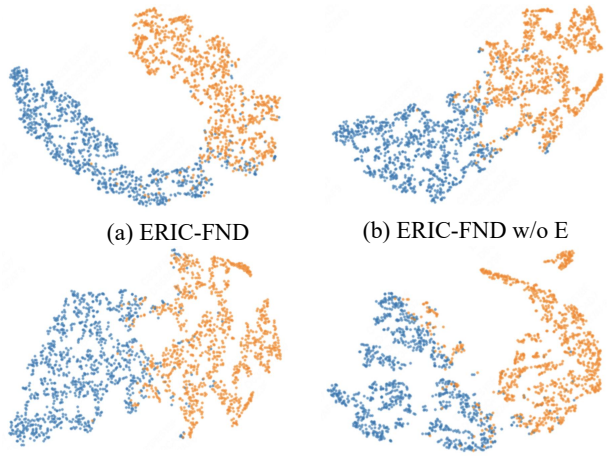
Experiment Environment The main parameters of the server computing environment are shown as follows:

CPU: AMD EPYC 7642 CPU @ 3293MHz 48C96T;
GPU: NVIDIA A100-SXM4-80GB;
CUDA Version: 11.0;
Memory: 32GB DDR4 2666MHz ECC ×16;
Operating System: CentOS Linux 7 (Core) Linux 3.10.0.
The running details are shown as follows:
Memory needed: 10917MiB;
Total number of model parameters: 132,350,153.

Parameter Settings and Implementation Details The settings of used parameters and the implementation details are described in the Appendix Section. The random seed is set and each experiment is run five times to ensure the reproduciblity.

Baselines We select several models in recent multimodal fake news detection studies as baseline models and the variants of proposed model ERIC-FND for comparison.

Recent multimodal fake news detection studies include **att-RNN** (Jin et al. 2017b), **EANN** (Wang et al. 2018), **MVAE** (Khattar et al. 2019b), **CAFE** (Chen et al. 2022b), **MCAN** (Wu et al. 2021), **MRML** (Peng et al. 2023) and **KMAGCN**(Qian et al. 2021).



(a) ERIC-FND (b) ERIC-FND w/o E

(c) ERIC-FND w/o M (d) ERIC-FND w/o A

Figure 3: The t-SNE feature visualization of proposed model ERIC-FND and its variants on Weibo dataset.

Variants of proposed model ERIC-FND:

ERIC-FND w/o A: The variant of proposed model ERIC-FND in this paper without adaptive fusion-based classification module. The final news representation is the simple concatenation of external information enhanced text features, the multimodal interaction enhancement feature and the representation of the news image.

ERIC-FND w/o M: The variant of proposed model ERIC-FND in this paper without multimodal information interactive enhancement module.

ERIC-FND w/o E: The variant of proposed model ERIC-FND in this paper without external information enhancement module.

All models are evaluated with the widely used metrics: accuracy, precision, recall and F1 score.

Comparative Experiment

The comparative experiment results are presented in Table 3 and the best results are in bold. Our proposed model ERIC-FND mostly outperforms the SOTA multimodal fake news detection models in accuracy, precision, recall and F1-score with different categories. Therefore, the comparative experiment results demonstrate that our proposed model ERIC-FND effectively exploits multimodal news information and the external information enhancement helps deepen the model understanding of textual semantics with obtained external reliable information. What is more, the adaptive fusion makes contributions to the selection of key features for detection.

Ablation Study

The results of ablation test are shown in Table 4. It can be seen from the table that the proposed model outperforms all the variants on both datasets. The removal of different module has various impact on the model performance. To be specific, after removing the adaptive fusion-based classification module, the model performance drops 0.8% accuracy

Dataset	Model	Accuracy	Fake News			Real News		
			Precision	Recall	F1-score	Precision	Recall	F1-score
Weibo	att-RNN	0.772	0.854	0.656	0.742	0.720	0.889	0.795
	EANN	0.782	0.827	0.697	0.756	0.752	0.863	0.804
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	CAFE	0.840	0.855	0.830	0.842	0.825	0.851	0.837
	MCAN	0.899	0.913	0.889	0.901	0.884	0.909	0.897
	MRML	0.897	0.898	0.887	0.892	0.896	0.905	0.901
	KMAGCN$_{bert}$	0.922	**0.993**	0.851	0.917	0.869	**0.994**	0.927
	ERIC-FND (proposed)	**0.946**	0.985	**0.914**	**0.948**	**0.908**	0.984	**0.944**
X	att-RNN	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN	0.648	0.810	0.498	0.617	0.584	0.759	0.660
	MVAE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	CAFE	0.806	0.807	0.799	0.803	0.805	0.813	0.809
	MCAN	0.809	0.889	0.765	0.822	0.732	0.871	0.795
	MRML	0.803	0.821	0.844	0.832	0.777	0.747	0.762
	KMAGCN$_{bert}$	0.804	0.787	0.784	0.785	0.817	0.819	0.818
	ERIC-FND (proposed)	**0.945**	**0.987**	**0.910**	**0.947**	**0.905**	**0.986**	**0.944**

Table 3: Comparative experiment results on Weibo dataset and X dataset.

Model	Accuracy	Precision	Recall	F1-score
ERIC-FND w/o A	0.938	0.939	0.938	0.938
ERIC-FND w/o M	0.925	0.925	0.932	0.925
ERIC-FND w/o E	0.926	0.933	0.926	0.926
ERIC-FND	**0.946**	**0.949**	**0.946**	**0.946**

Table 4: Ablation test results on Weibo dataset. A represents adaptive fusion-based classification module. M represents multi-modal information interactive enhancement module. E represents external information enhancement module.

on both datasets, which indicates the adaptive fusion is able to help extract key features among all the obtained feature representation. With the multimodal information interactive enhancement module, the model separately achieves an improvement of 2.1% and 1.6% in Weibo and X dataset compared to the variant model, which proves that the multimodal information interactive enhancement enables to capture the deep semantic interaction of multimodal information. As to the addition of external information enhancement module, there are increases of 2% and 3.1% in accuracy on Weibo and X datasets which shows that extracted entity description helps enhance the news textual representation.

Visualization Analysis

The visualization analysis of the obtained news representations before being input into the classifier is conducted to further validate the effectiveness of our proposed model. Specifically, we apply the t-distributed Stochastic Neighbor Embedding (t-SNE) method (van der Maaten and Hinton 2008) on the Weibo dataset to perform dimensionality reduction on the obtained news representations before being input into the classifier. The dimensionality-reduced data are then visualized which provides a clear observation of the distribution of news data points with different labels in low dimensionality space. Figure 3 shows the visualization results of proposed model ERIC-FND and its variant models on Weibo dataset. In the figure, blue points represent real news, and orange points represent fake news.

As shown in the figure, the news feature representations obtained by ERIC-FND are more distinct with different labels in the low dimensionality, which indicates that the proposed model has the capability of effectively differentiating different categories of news pieces. This result demonstrates that the proposed model is able to largely enhance the final news representation features, further proving the model effectiveness.

Conclusion

In this paper, we propose the ERIC-FND model, which employs entity-based external reliable information to enhance the understanding of textual content and multimodal information interactive semantic enhancement to improve cross-modality learning. Finally, an adaptive fusion method is utilized to optimize the contribution of each news feature to the final classification. Experimental results demonstrate that ERIC-FND outperforms SOTA models and its variants on two datasets, indicating the effectiveness of the ERIC-FND architecture.

Acknowledgements

This work is supported by National Natural Science Foundation of China under Grants No.62472092, No.62172089, No.62172090, No.62106045. Natural Science Foundation of Jiangsu province under Grants No.BK20241751. Jiangsu Provincial Key Laboratory of Computer Networking Technology. Jiangsu Provincial Key Laboratory of Network and Information Security under Grants No.BM2003201, and Key Laboratory of Computer Network and Information Integration of Ministry of Education of China under Grants No.93K-9, Nanjing Purple Mountain Laboratories. Startup Research Fund of Southeast University under Grants No.RF1028623097. We thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations.

References

Boididou, C.; Andreadou, K.; Papadopoulos, S.; Dang Nguyen, D. T.; Boato, G.; Riegler, M.; Larson, M.; and Kompatsiaris, I. 2015. Verifying Multimedia Use at MediaEval 2015 In MediaEval Benchmarking Initiative for Multimedia Evaluation. In *MediaEval2015*.

Chen, T.; and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 785–794. New York, NY, USA: Association for Computing Machinery. ISBN 9781450342322.

Chen, Y.; Li, D.; Zhang, P.; Sui, J.; Lv, Q.; Tun, L.; and Shang, L. 2022a. Cross-modal Ambiguity Learning for Multimodal Fake News Detection. In *Proceedings of the ACM Web Conference 2022*, WWW '22, 2897–2905. New York, NY, USA: Association for Computing Machinery. ISBN 9781450390965.

Chen, Y.; Li, D.; Zhang, P.; Sui, J.; Lv, Q.; Tun, L.; and Shang, L. 2022b. Cross-modal Ambiguity Learning for Multimodal Fake News Detection. In *Proceedings of the ACM Web Conference 2022*, WWW '22, 2897–2905. New York, NY, USA: Association for Computing Machinery. ISBN 9781450390965.

Fang, L.; Feng, K.; Zhao, K.; Hu, A.; and Li, T. 2023. Unsupervised Rumor Detection Based on Propagation Tree VAE. *IEEE Transactions on Knowledge and Data Engineering*, 35(10): 10309–10323.

Feng, S.; Banerjee, R.; and Choi, Y. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, 171–175. USA: Association for Computational Linguistics.

Gupta, A.; Lamba, H.; Kumaraguru, P.; and Joshi, A. 2013. Faking Sandy: characterizing and identifying fake images on Twitter during Hurricane Sandy. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion, 729–736. New York, NY, USA: Association for Computing Machinery. ISBN 9781450320382.

Hu, L.; Yang, T.; Zhang, L.; Zhong, W.; Tang, D.; Shi, C.; Duan, N.; and Zhou, M. 2021. Compare to The Knowledge: Graph Neural Fake News Detection with External Knowledge. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 754–763. Online: Association for Computational Linguistics.

Hu, X.; Guo, Z.; Chen, J.; Wen, L.; and Yu, P. S. 2023. MR2: A Benchmark for Multimodal Retrieval-Augmented Rumor Detection in Social Media. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, 2901–2912. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394086.

Jin, Z.; Cao, J.; Guo, H.; Zhang, Y.; and Luo, J. 2017a. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, 795–816. New York, NY, USA: Association for Computing Machinery. ISBN 9781450349062.

Jin, Z.; Cao, J.; Guo, H.; Zhang, Y.; and Luo, J. 2017b. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, 795–816. New York, NY, USA: Association for Computing Machinery. ISBN 9781450349062.

Jin, Z.; Cao, J.; Zhang, Y.; and Luo, J. 2016. News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, 2972–2978. AAAI Press.

Jin, Z.; Cao, J.; Zhang, Y.; Zhou, J.; and Tian, Q. 2017c. Novel Visual and Statistical Image Features for Microblogs News Verification. *Trans. Multi.*, 19(3): 598–608.

Khattar, D.; Goud, J. S.; Gupta, M.; and Varma, V. 2019a. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In *The World Wide Web Conference*, WWW '19, 2915–2921. New York, NY, USA: Association for Computing Machinery. ISBN 9781450366748.

Khattar, D.; Goud, J. S.; Gupta, M.; and Varma, V. 2019b. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In *The World Wide Web Conference*, WWW '19, 2915–2921. New York, NY, USA: Association for Computing Machinery. ISBN 9781450366748.

Lin, H.; Yi, P.; Ma, J.; Jiang, H.; Luo, Z.; Shi, S.; and Liu, R. 2023. Zero-shot rumor detection with propagation structure via prompt learning. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press. ISBN 978-1-57735-880-0.

Liu, Y.; and Wu, Y.-F. B. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the Thirty-Second AAAI Conference on Artifi-*

cial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18. AAAI Press. ISBN 978-1-57735-800-8.

Ma, J.; Gao, W.; Mitra, P.; Kwon, S.; Jansen, B. J.; Wong, K.-F.; and Cha, M. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, 3818–3824. AAAI Press. ISBN 9781577357704.

Ma, J.; Gao, W.; and Wong, K.-F. 2017. Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning. In Barzilay, R.; and Kan, M.-Y., eds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 708–717. Vancouver, Canada: Association for Computational Linguistics.

Ma, J.; Gao, W.; and Wong, K.-F. 2019. Detect Rumors on Twitter by Promoting Information Campaigns with Generative Adversarial Learning. In *The World Wide Web Conference*, WWW '19, 3049–3055. New York, NY, USA: Association for Computing Machinery. ISBN 9781450366748.

Peng, L.; Jian, S.; Li, D.; and Shen, S. 2023. MRML: Multimodal Rumor Detection by Deep Metric Learning. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Pérez-Rosas, V.; Kleinberg, B.; Lefevre, A.; and Mihalcea, R. 2018. Automatic Detection of Fake News. In Bender, E. M.; Derczynski, L.; and Isabelle, P., eds., *Proceedings of the 27th International Conference on Computational Linguistics*, 3391–3401. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Qi, P.; Cao, J.; Yang, T.; Guo, J.; and Li, J. 2019. Exploiting Multi-domain Visual Information for Fake News Detection. In *2019 IEEE International Conference on Data Mining (ICDM)*, 518–527.

Qian, S.; Hu, J.; Fang, Q.; and Xu, C. 2021. Knowledge-aware Multi-modal Adaptive Graph Convolutional Networks for Fake News Detection. *ACM Trans. Multimedia Comput. Commun. Appl.*, 17(3).

Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017a. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newsl.*, 19(1): 22–36.

Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017b. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newsl.*, 19(1): 22–36.

Sun, X.; Yin, H.; Liu, B.; Meng, Q.; Cao, J.; Zhou, A.; and Chen, H. 2023. Structure Learning Via Meta-Hyperedge for Dynamic Rumor Detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(9): 9128–9139.

Tseng, Y.-W.; Yang, H.-K.; Wang, W.-Y.; and Peng, W.-C. 2022. KAHAN: Knowledge-Aware Hierarchical Attention Network for Fake News detection on Social Media. In *Companion Proceedings of the Web Conference 2022*, WWW '22, 868–875. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391306.

van der Maaten, L.; and Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86): 2579–2605.

Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; and 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*, KDD '18, 849–857. New York, NY, USA: Association for Computing Machinery. ISBN 9781450355520.

Wu, Y.; Zhan, P.; Zhang, Y.; Wang, L.; and Xu, Z. 2021. Multimodal Fusion with Co-Attention Networks for Fake News Detection. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2560–2569. Online: Association for Computational Linguistics.

Yu, F.; Liu, Q.; Wu, S.; Wang, L.; and Tan, T. 2017. A convolutional approach for misinformation identification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, 3901–3907. AAAI Press. ISBN 9780999241103.

Zhang, L.; Zhang, X.; Zhou, Z.; Huang, F.; and Li, C. 2024. Reinforced Adaptive Knowledge Learning for Multimodal Fake News Detection. In Wooldridge, M. J.; Dy, J. G.; and Natarajan, S., eds., *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, 16777–16785. AAAI Press.

Zhou, X.; Jain, A.; Phoha, V. V.; and Zafarani, R. 2020. Fake News Early Detection: A Theory-driven Model. *Digital Threats*, 1(2).

Zhou, X.; and Zafarani, R. 2019. Network-based Fake News Detection: A Pattern-driven Approach. *SIGKDD Explor. Newsl.*, 21(2): 48–60.