Information Processing and Management 62 (2025) 104089

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

DCCMA-Net: Disentanglement-based cross-modal clues mining and aggregation network for explainable multimodal fake news detection

Siqi Wei, Zheng Wang, Meiling Li, Xuanning Liu, Bin Wu *

School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China

ARTICLE INFO

Keywords:
Multimodal fake news detection
Disentangled representation learning
Cross-modal clues exploration
Explainable machine learning

ABSTRACT

Multimodal fake news detection is significant in safeguarding social security. Compared with single-text news, multimodal news data contains rich cross-modal clues that can improve the detection effectiveness: modality-common semantic enhancement, modality-specific semantic complementation, and modality-specific semantic inconsistency. However, most existing studies ignore the disentanglement of modality-specific and modality-common semantics but treat them as an entangled whole. Consequently, these studies can only implicitly explore the interactions between modalities, resulting in a lack of explainability. To address that, we propose a Disentanglement-based Cross-modal Clues Mining and Aggregation Network for explainable fake news detection, called DCCMA-Net. Specifically, DCCMA-Net decomposes each modality into two distinct representations: a modality-common representation that captures shared semantics across modalities, and a modality-specific representation that captures unique semantics within each modality. Then, leveraging these disentangled representations, DCCMA-Net explicitly and comprehensively mines three cross-modal clues: modality-common semantic enhancement, modality-specific semantic complementation, and modality-specific semantic inconsistency. Since not all clues play an equal role in the decision-making process, DCCMA-Net proposes an adaptive attention aggregation module to assign contribution weights to different clues. Finally, DCCMA-Net aggregates these clues based on their contribution weights to obtain highly discriminative news representations for detection, and highlights the most contributive clues as explanations for the detection results. Extensive experiments demonstrate that DCCMA-Net outperforms existing methods, achieving detection accuracy improvements of 2.53%, 4.01%, and 3.99% on Weibo, PHEME, and Gossipcop datasets, respectively. Moreover, the explainability accuracy of DCCMA-Net exceeds that of current state-of-the-art methods on the Weibo dataset.

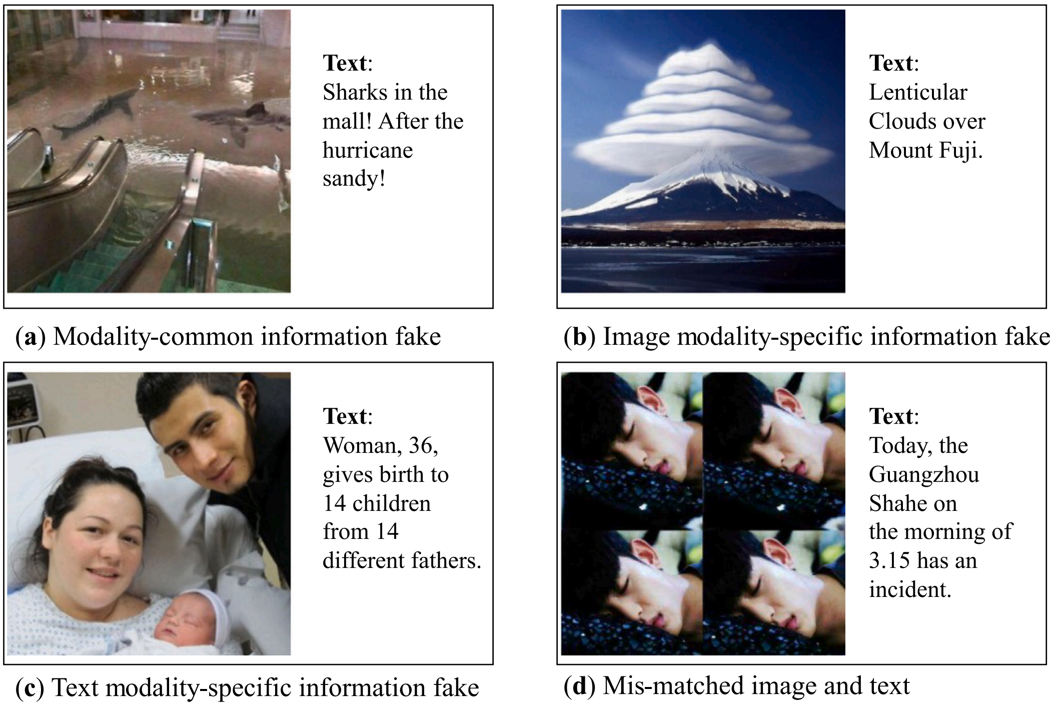
1. Introduction

The rapid development of social networks greatly facilitates people to obtain and share information. However, the lack of regulation also leads to the widespread propagation of fake news, which causes serious harm to both individuals and society (Zhang & Ghorbani, 2020). In order to mitigate the negative effects caused by fake news, some fact-checking organizations, such as Politifact[1]

* Corresponding author.
E-mail addresses: Weisiqi@bupt.edu.cn (S. Wei), Wangzheng123@bupt.edu.cn (Z. Wang), meilinglee@bupt.edu.cn (M. Li), liuxuanning@bupt.edu.cn (X. Liu), wubin@bupt.edu.cn (B. Wu).

[1] https://www.politifact.com.

https://doi.org/10.1016/j.ipm.2025.104089
Received 29 May 2024; Received in revised form 29 December 2024; Accepted 29 January 2025
Available online 14 February 2025



Text:
Sharks in the mall! After the hurricane sandy!

(**a**) Modality-common information fake

Text:
Lenticular Clouds over Mount Fuji.

(**b**) Image modality-specific information fake

Text:
Woman, 36, gives birth to 14 children from 14 different fathers.

(**c**) Text modality-specific information fake

Text:
Today, the Guangzhou Shahe on the morning of 3.15 has an incident.

(**d**) Mis-matched image and text

Fig. 1. Some examples of multimodal fake news. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and FactCheck,[2] manually detect fake news (Zhou, Zafarani, Shu, & Liu, 2019). Nevertheless, manual detection requires a large number of well-trained experts in the journalism field, which is time-consuming and labor-intensive. Additionally, the massive amount of fake news on social platforms makes relying entirely on manual detection impossible. Therefore, automatic fake news detection has become a critical issue that needs to be solved urgently (Bondielli & Marcelloni, 2019).

The objective of automatic fake news detection is to identify the veracity of news articles through the analysis of their features (Shu, Sliva, Wang, Tang, & Liu, 2017). Existing studies in this field can be broadly classified into two categories based on the number of modalities: unimodal and multimodal approaches (Hu, Wei, Zhao, & Wu, 2022). Unimodal methods capture news text features for detection. News text content, as the main component of news, is the main basis for detecting fake news (Zhou et al., 2019). Earlier methods (Castillo, Mendoza, & Poblete, 2011; Popat, Mukherjee, Strötgen, & Weikum, 2016) manually extracted linguistic features and input them into shallow machine learning models to obtain classification results. However, these machine learning-based methods rely heavily on complex feature engineering and lack generalization (Hu et al., 2022). With the development of deep learning, a large number of studies utilize deep neural networks, such as the Recurrent Neural Network (RNN) (Ma et al., 2016), transformer (Wani, Joshi, Khandve, Wagh, & Joshi, 2021), or GPT-3.5 (Hu et al., 2024) to automatically mine news semantics. Moreover, in addition to news text content, some studies capture additional text from news comments (Luvembe, Li, Li, Liu, & Xu, 2023; Shu, Cui, Wang, Lee, & Liu, 2019) or external knowledge (Guo, Zeng, Tang, & Zhao, 2023; Jiang, Liu, Zhao, Sun, & Zhang, 2022; Zhang et al., 2021) to enhance detection accuracy. Despite achieving some results, the availability of such additional texts is not always guaranteed. Nevertheless, as multimedia technology progresses, news articles are evolving from text form to multimodal form that include both images and text (Alam, Cresci, et al., 2022). Consequently, unimodal methods are inadequate for effectively analyzing multimodal news content.

In contrast to single-text news, multimodal news incorporates rich semantic associations between images and text, which can provide abundant evidence for detecting fake news. We mine the following three meaningfully cross-modal clues: (1) modality-common semantic enhancement. The image and text jointly describe the same news event, so they convey some shared semantics called modality-common semantics. The common parts often contain important elements of the event. In the news shown in Fig. 1(a), both the image and the text describe the shark and the shopping mall, which are important elements of the news event. Although these modality-common semantics do not provide additional information, different modalities can provide different views and enhance the model's understanding of the important elements of the news. Specifically, the image provides pattern and color information in the visual view; the text provides entity and context information in the linguistic view. (2) Modality-specific semantic complementarity. Each modality contains its private semantics, called modality-specific semantics. They complement each

[2] https://www.factcheck.org.

other and contain discriminative features. In the news shown in Fig. 1(b) and (c), the image-specific and text-specific semantics provide evidence for the detection results, respectively. (3) Modality-specific semantic inconsistency. Because it is difficult to find semantically consistent real images to support non-factual content in fake news, incorrectly using irrelevant images is a common way to fabricate fake news. In the news shown in Fig. 1(d), the Korean star presented in the image is semantically inconsistent with the terrorist event presented in the text. In this type of fake news, the inconsistency between modality-specific semantics provides evidence for detection.

However, most exciting studies (Khattar, Goud, Gupta, & Varma, 2019; Singhal et al., 2020; Wang et al., 2018, 2023) utilize an integrated representation to represent each modality, ignoring the disentanglement between modality-common and modality-specific semantics. These entangled representations lead to models only implicitly mining cross-modal clues. Therefore, they are at the black-box level, which means that they cannot explicitly provide evidence to explain the detection results. Explainable Fake News Detection aims to accurately identify fake news and provide evidence to explain the detection results. Explainability is essential for building public trust and effectively debunking fake news (Reis, Correia, Murai, Veloso, & Benevenuto, 2019). However, the disentanglement of modality-specific and modality-common semantics has not been thoroughly investigated, and thus explicit mining of cross-modal clues cannot be guaranteed, which leads to the inability to provide evidence for detection results. Therefore, how to comprehensively and explicitly mine multiple cross-modal clues is an important challenge in addressing explainable multimodal fake news detection.

In addition, how different clues affect decision-making is another issue in explainable fake news detection. Different cross-modal clues contribute differently to detecting different fake news. For instance, the modality-common semantic enhancement clue is most influential in detecting the fake news depicted in Fig. 1(a). Conversely, the modality-specific semantic complementarity clue plays a primary role in identifying the fake news illustrated in Fig. 1(b) and (c). The modality-specific semantic inconsistency clue is most effective in detecting the fake news presented in Fig. 1(d). Consequently, how to adaptively aggregate different cross-modal clues and highlight the most contributive clue is another important challenge for explainable fake news detection.

To address the above issues, we propose a Disentanglement-based Cross-modal Clues Mining and Aggregation Network called DCCMA-Net, for explainable multimodal fake news detection. This framework contains five main components. (1) Modality feature extraction module (MFE) employs a pre-trained CLIP model (Radford et al., 2021) to extract image and text features and map them to a unified space to mitigate modality heterogeneity. (2) Multimodal disentanglement autoencoder module (MDAE) disentangles each modality into two refined representations to mitigate modality redundancy. The first representation is modality-common, which focuses on capturing shared semantics across modalities. The second representation is modality-specific, which aims to capture unique semantics specific to each modality. (3) Cross-modal clues mining module (CCM) utilizes the disentangled representations to explore three cross-modal clues comprehensively. First, CCM concatenates text-specific and image-specific representations to model the modality-specific semantic complementation. Second, CCM designs a cross-modal mutual enhancement unit to model the modality-common semantics enhancement. Third, CCM proposes a cross-modal inconsistency comparison unit to model the modality-specific semantic inconsistency. (4) Adaptive attention aggregation module (AAA) reweights and aggregates different clues to obtain the highly discriminative news representation and emphasizes the most contributing clue to explain the detection result. (5)Classification module determines the veracity of the news based on its representation.

We conduct experiments on three public multimodal fake news detection datasets: Weibo (Jin, Cao, Guo, Zhang, & Luo, 2017), PHEME (Zubiaga, Liakata, & Procter, 2017), and Gossipcop (Shu, Mahudeswaran, Wang, Lee, & Liu, 2020). Experimental results show that DCCMA-Net outperforms existing state-of-the-art multimodal fake news detection methods in terms of performance and explainability. Furthermore, we fully discuss the effectiveness of each innovation strategy of our model. Finally, we visually demonstrate the capability of DCCMA-Net in mining and aggregating cross-modal clues by analyzing typical cases. The main contributions of this paper can be summarized as follows:

- We emphasize the explainability of multimodal fake news detection task and propose a novel DCCMA-Net model to effectively detect fake news and provide explanations for the detection results.
- DCCMA-Net mines multiple cross-modal clues explicitly and comprehensively to provide sufficient evidence for detecting fake news.
- DCCMA-Net adaptively aggregates cross-modal clues and highlights the most contributive clues to explain the detection results.

- Experiment results on three public datasets demonstrate the effectiveness and explainability of our proposed DCCMA-Net.

The organization of this paper's succeeding sections is as follows. In Section 2, we review existing studies in fake news detection. In Section 3, we introduce the research objective of explainable multimodal fake news detection. In Section 4, we describe the design detail of our proposed DCCMA-Net model. In Section 5, we perform extensive experiments and analyze the results. In Section 6, we discuss the theoretical and practical implications of DCCMA-Net as well as our future research directions. Finally, in Section 7, we provide a conclusion of the whole paper.

2. Related work

2.1. Unimodal method

Textual content is the main component of news, so a large number of researchers in the field of natural language processing (NLP) mine textual features to classify news authenticity. With the development of NLP technology, unimodal fake news detection has gone through the following stages: (1) Machine learning stage: Early research utilized manually extracted linguistic statistical features and fed them into shallow machine learning models (e.g., support vector machines Yang, Liu, Yu, & Yang, 2012, decision trees Castillo et al., 2011, and random forests Kwon, Cha, Jung, Chen, & Wang, 2013) to obtain classification results. However, machine learning-based approaches not only fail to capture news semantic information but also lack generalization. (2) End-to-end neural network stage: with the rise of deep learning, some end-to-end neural networks such as recurrent neural network (RNN) (Ma et al., 2016), Convolutional neural networks (CNN) (Yu et al., 2017), and graph neural networks (GNN) (Vaibhav & Hovy, 2019) are used to automatically study semantic features for detection. (3) Pre-train and fine-tune stage: Benefiting from the strong semantic understanding of pre-trained language models, some studies utilize fake news detection data to fine-tune pre-trained models for detection. For example, Kaliyar, Goswami, and Narang (2021) fine-tune BERT and Samadi, Mousavian, and Momtazi (2021) fine-tune Roberta, which both achieved good results. (4) Large language models and prompt learning stage: Large language models (LLMs) (Wang et al., 2024) contain rich semantic knowledge by pre-training on large corpus. Therefore, LLMs can achieve good results even without training. Some studies (Hu et al., 2024; Wang et al., 2024) utilize prompt learning techniques (Liu et al., 2023) to stimulate the potential of LLMs for detecting fake news with zero or few samples.

2.2. Multimodal method

The multimodal FND methods combine multiple modality information to detect fake news. Based on the modality fusion approach, existing methods can be broadly categorized into joint representation methods, alignment representation methods, and comparison representation methods.

Joint representation methods fuse image and text features by vector concatenation. For example, SpotFake (Singhal, Shah, Chakraborty, Kumaraguru, & Satoh, 2019) utilizes BERT (Kenton & Toutanova, 2019) to extract text features and VGG (Simonyan & Zisserman, 2015) to extract image features, followed by concatenating them to obtain the multimodal news representation. Similarly, SpotFake+ (Singhal et al., 2020) concatenates image and text features extracted by XLNet (Yang et al., 2019) and VGG. Based on this framework, some studies add auxiliary tasks to help the model better understand the multimodal features of news. For example, EANN (Wang et al., 2018) proposes an event classification auxiliary task, and MVAE (Khattar et al., 2019) proposes a modality feature reconstruction auxiliary task. In addition, Wei et al. (2022) propose a cross-modal knowledge distillation approach to enhance the unimodal encoder and then aggregate image and text features for classification. However, existing studies tend to treat image and text features as equally significant, overlooking the distinct contributions of each modality to the detection results. To address this limitation, Singhal, Pandey, Mrig, Shah, and Kumaraguru (2022) proposed a multiplicative fusion method to combine different modality features based on their respective contributions. Despite the simplicity and effectiveness of joint representation methods, they only consider images as a complement to text, ignoring the rich semantic correlations between modalities. Moreover, they also ignore modality redundancy.

The news text and its corresponding image express the same news event. Therefore, the modality-common semantics always contain the critical elements of the news. The **alignment representation methods** fuse the image and text by aligning their features. Jin et al. (2017) use an attention-based RNN to highlight the text token associated with the image. Inspired by the self-attention mechanism in transformer (Vaswani et al., 2017), some studies (Qian, Wang, Hu, Fang, & Xu, 2021; Song, Ning, Zhang, & Wu, 2021; Wu, Zhan, Zhang, Wang, & Xu, 2021; Zhang, Gui and He, 2021) utilize modified co-attention transformer to align the features form image and text. Jing, Wu, Sun, Fang, and Zhang (2023) propose a progressive fusion network to align higher-order and lower-order multimodal features simultaneously. Zhou, Yang, Ying, Qian, and Zhang (2023a) propose a multi-grained multimodal fusion network to achieve coarse-grained and fine-grained mutual enhancement between different modalities. However, modality-specific semantics are usually unrelated to other modalities and are treated as noise during the alignment process.

The phenomenon of mismatched image–text pairs often occurs in fake news. Hence, **comparison representation approaches** compare the inconsistency between the image and the text for detection. Zhou, Wu, and Zafarani (2020) transform the image into caption text by utilizing an image caption model and measure the consistency between the news text and the caption text. Xue et al. (2021) map the image and the text into the same embedding space and calculate their cosine similarity. Chen et al. (2022) measure the modality ambiguity by calculating the KL divergence between different modalities. Shang, Kou, Zhang, and Wang (2022) design a duo-generative cross-modal feature generation approach to generate text-guided visual and image-guided textual features simultaneously and measure the inconsistency between the original and generated features to detect misinformation. Zhou et al. (2023a) use the CLIP similarity score to measure the inconsistency between modalities. Wang et al. (2023) propose a cross-modal contrastive learning framework to project images and text into a unified space and measure their inconsistency. However, the modality-common semantics emphasizes the commonality and affects the comparison process.

The **hybrid approaches** aim to mine multiple cross-modal clues to provide more abundant evidence for fake news detection. EM-FEND (Qi et al., 2021) explicitly extracts the visual entities and models the multimodal inconsistency and enhancement based on the multimodal entities. Moreover, EM-FEND extracts the embedded text in the images and models the text complementation. BMR (Ying et al., 2023) extracts features respectively from the views of the text, the image pattern, and the image semantics. Then BMR model inconsistency and complementation relationships between these features. Although they achieve improved performance, they still cannot consider the enhancement, complementation, inconsistency, and redundancy between modalities simultaneously.

To summarize, existing studies are unable to distinguish between modality-common and modality-specific semantics, and instead treat each modality as an entangled whole. These entangled representations lead to models that cannot explicitly mine cross-modal clues. Therefore most existing methods are at the black-box level and cannot provide explanations for detection. In addition, existing methods are unable to highlight the most contributive clues. Unlike existing approaches, our proposed DCCMA-Net disentangles each modality into refined modality-common and modality-specific representations and utilizes these disentangled representations to explicitly and comprehensively mine cross-modal clues. Specifically, DCCMA-Net first models the enhancement correlations between modality-specific semantics to avoid the noise introduced by modality-specific semantics during the alignment process. DCCMA-Net compares inconsistencies between modality-specific semantics, avoiding the effect of modality-common semantics during the inconsistency comparison process. DCCMA-Net models complementary relationships between modality-specific semantics, mitigating modal redundancy. Finally, an adaptive attention aggregation mechanism is proposed for highlighting the most contributing clues.

3. Research objective

In this section, we formally present the research objective of explainable multimodal fake news detection. We first define some key terms that will be used in the research objective statement.

Definition 1 (*Multimodal News N*)**.** Multimodal news refers to online news articles that contain both text and images. Specifically, given a piece of multimodal news as $N = \{(T, V), y\}$, where T denotes the text content, V denotes the image content and $y \in \{0, 1\}$ denotes the ground-truth label of the news ((i.e., real or fake).

Definition 2 (*Modality-Common and Modality-Specific Semantics $h^c \& h^s$*)**.** Given a piece of multimodal fake news N, each modality (e.g., image V and text I) contains two types of semantics: modality-common and modality-specific.

- Modality-common semantics ($\mathbf{h}^c$) refer to the consistent semantics between different modalities. Since different modalities represent the same news event, shared semantics exist.
- Modality-specific semantics ($\mathbf{h}^s$) refer to the semantics unique to each modality, which is different from other modalities.

Definition 3 (*Cross-Modal Clues*)**.** Cross-modal clues refer to rich semantic associations between different modalities, which can provide evidence for fake news detection. We summarize three types of cross-modal clues: modality-common semantic enhancement, modality-specific semantic complementation, and modality-specific semantic inconsistency. Detailed explanations are as follows:

- **Modality-common semantic enhancement** ($\mathbf{h}_{me}$): Modality-common semantics do not provide additional information, but different modalities can enhance each other and enable the model to understand the news better. We denote this association as modality-common semantic enhancement.
- **Modality-specific semantic complementation** ($\left[\mathbf{h}_t^s, \mathbf{h}_v^s\right]$): Specific semantics from the image and text can complement each other and describe the complete news event together. We denote this association as modality-specific information Complementation.
- **Modality-specific semantic inconsistency** ($\mathbf{h}_{mi}$): Incorrectly using mismatched image–text is a common way to create fake news. Comparing differences between modality-specific semantics can highlight the inconsistency between image and text even more.

Definition 4 (*Fake News Detection Explanation*)**.** The explanation refers to the different cross-modal clues and the weight of their contribution to the detection results. Formally, we define the explanation as $w = \left\{w_v, w_t, w_{me}, w_{mi}\right\}$, where $w_v, w_t, w_{me}, w_{mi} \in [0, 1]$ denote the contribution weights of image-specific semantic, text-specific semantic, modality-common semantic enhancement, and modality-specific semantic inconsistency to the model decision, respectively.

Problem statement. In summary, the goal of our study is to detect the authenticity of multimodal news and provide an explanation for the detection result. Specifically, given a multimodal news $N = (T, V)$, we first decompose each modality into modality-common semantics ($\mathbf{h}_t^c$ or $\mathbf{h}_v^c$) and modality-specific semantics ($\mathbf{h}_t^s$ or $\mathbf{h}_v^s$). Following this, the disentangled representations are used to explicitly mine cross-modal clues as $\left[\mathbf{h}_t^s, \mathbf{h}_v^s, \mathbf{h}_{me}, \mathbf{h}_{mi}\right]$. After that, we assign weights to each representation based on its contribution to news veracity as $\left[w_t^s, w_v^s, w_{me}, w_{mi}\right]$. Finally, we aggregate these clues based on the contribution weights to obtain the highly discriminative news representation $\mathbf{x}$ to detect fake news and highlight the most contributive clues as explanations for the detection results.

4. Methodology

4.1. Overview

To improve the effectiveness and explainability of multimodal fake news detection, we propose a novel Disentanglement-based Cross-modal Clues Mining and Aggregation Network, named DCCMA-Net for short. Fig. 2 depicts the overall model structure. DCCMA-Net mainly consists of five components: (1) Modality feature extraction module (MFE), which extracts text and image features and maps them into a unified embedding space; (2) Multimodal disentanglement autoencoder module (MDAE), which decomposes each modality into refined modality-common and modality-specific representations to mitigate redundancy; (3) Cross-modal clues mining module (CCM), which comprehensively and explicitly explores various cross-modal clues based on disentangled representations. (4) Adaptive attention aggregation module (AAA), which reweights and aggregates different clues to boost performance and provide explainability; and 5) Classification module, which outputs detection results.

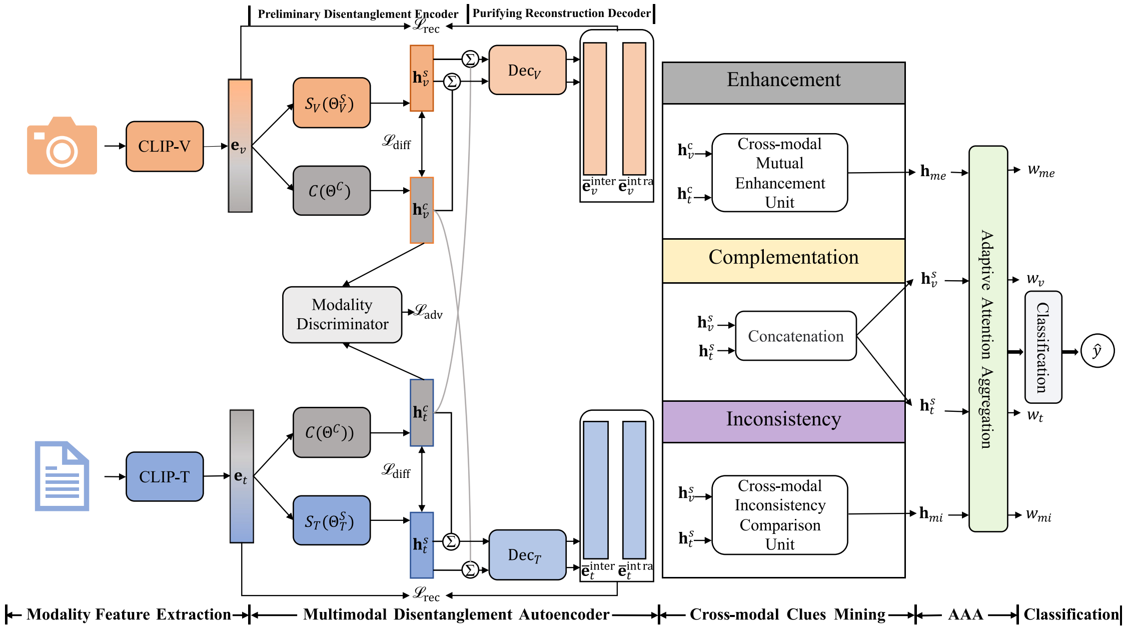

Preliminary Disentanglement Encoder
Purifying Reconstruction Decoder
$\mathcal{L}_{\text{rec}}$
CLIP-V
$\mathbf{e}_v$
$S_V(\Theta_V^S)$
$C(\Theta^C)$
$\mathbf{h}_v^s$
$\mathbf{h}_v^c$
$\mathcal{L}_{\text{diff}}$
Σ
Dec_V
$\bar{\mathbf{e}}_v^{\text{inter}}$
$\bar{\mathbf{e}}_v^{\text{intra}}$
Modality Discriminator
$\mathcal{L}_{\text{adv}}$
CLIP-T
$\mathbf{e}_t$
$C(\Theta^C)$
$S_T(\Theta_T^S)$
$\mathbf{h}_t^c$
$\mathbf{h}_t^s$
Dec_T
$\bar{\mathbf{e}}_t^{\text{inter}}$
$\bar{\mathbf{e}}_t^{\text{intra}}$
Enhancement
Cross-modal Mutual Enhancement Unit
$\mathbf{h}_{me}$
Complementation
Concatenation
Inconsistency
Cross-modal Inconsistency Comparison Unit
$\mathbf{h}_{mi}$
Adaptive Attention Aggregation
w_{me}
w_v
w_t
w_{mi}
Classification
$\hat{y}$
Modality Feature Extraction
Multimodal Disentanglement Autoencoder
Cross-modal Clues Mining
AAA
Classification

Fig. 2. The architecture of DCCMA-Net model.

4.2. Modality feature extraction

The modality feature extraction module is designed to mine image and text features of news and map them into a unified embedding space. CLIP (Radford et al., 2021), as an advanced multimodal pre-training model, can effectively capture rich multimodal features in news. In addition, benefiting from the contrastive learning mechanism, CLIP can map the rich semantic features of different modalities into a unified embedding space. Therefore, we employ a pre-trained CLIP model (Radford et al., 2021) to extract the features from text and image modality and map them into a unified embedding space as below:

$$\begin{cases} \mathbf{e}_t = \text{CLIP-T}(T), \\ \mathbf{e}_v = \text{CLIP-V}(V), \end{cases} \tag{1}$$

where $\mathbf{e}_t, \mathbf{e}_v \in \mathbb{R}^{d_c}$ denote the text and image representations of the news, respectively, and d_c indicates the vector dimension of the CLIP model output.

Previous studies (Singhal et al., 2019; Wang et al., 2023) always employ two unimodal encoders to project image and text features into distinct embedding spaces, which leads to a heterogeneity gap between modalities. In contrast, the CLIP model leverages the contrastive learning mechanism (Chen, Kornblith, Norouzi, & Hinton, 2020) to map different modality features into a unified embedding space, mitigating the problem of multimodal feature heterogeneity.

4.3. Multimodal disentanglement autoencoder

In order to obtain refined modality-common and modality-specific representations, we propose a multimodal disentanglement autoencoder module, which connects a preliminary disentanglement encoder and a purifying reconstruction decoder in an end-to-end manner. The preliminary disentanglement encoder first roughly maps the features of each modality into modality-common and modality-specific subspaces. Then the purifying reconstruction decoder refines both subspaces, which can disentangle deeply to mitigate redundancy.

4.3.1. Preliminary disentanglement encoder

Given the vectors for each modality, we propose the common and specific project functions to map each modality into modality-common and modality-specific subspaces, respectively. Following the classical modality disentanglement representation learning work (Yang, Huang, Kuang, Du, & Zhang, 2022), Both the common project function C and the specific project functions S_T, S_V are implemented as two-layer perceptions (MLP) with the GeLU activation function (Hendrycks & Gimpel, 2016). The common and specific project functions can be formulated as:

$$\begin{cases} \mathbf{h}_t^c = C\left(\mathbf{e}_t; \Theta^C\right), \\ \mathbf{h}_v^c = C\left(\mathbf{e}_v; \Theta^C\right), \\ \mathbf{h}_t^s = S_T\left(\mathbf{e}_t; \Theta_T^S\right), \\ \mathbf{h}_v^s = S_V\left(\mathbf{e}_V; \Theta_V^S\right), \end{cases} \tag{2}$$

where $\mathbf{h}_t^c, \mathbf{h}_v^c, \mathbf{h}_t^s, \mathbf{h}_v^s \in \mathbb{R}^d$. The common project function $C\left(\cdot, \Theta^C\right)$ shares the parameters Θ^C across two modalities, whereas the specific project functions $S_V\left(\cdot, \Theta_v^S\right)$ and $S_T\left(\cdot, \Theta_T^S\right)$ assign separate parameters Θ_T^S and Θ_V^S for each modality.

Based on these aforementioned common and specific project functions, we introduce the difference and adversarial similarity constraints to ensure the disentanglement of common and specific semantics. The difference constraint ensures the distinction between modality-common and modality-specific semantics. Meanwhile, the adversarial similarity constraint employs adversarial training to ensure the commonality of modality-common semantics.

Difference Constraint: The difference constraint ensures that modality-common and modality-specific representations can capture different aspects of each modality. Moreover, the difference between the two can effectively ensure their non-redundancy. Specifically, the difference constraint can be defined as follows:

$$\mathcal{L}_{\text{diff}} = \cos\left(\mathbf{h}_t^c, \mathbf{h}_t^s\right) + \cos\left(\mathbf{h}_v^c, \mathbf{h}_v^s\right), \tag{3}$$

where $\cos(\cdot, \cdot)$ denotes the cosine similarity between two given input embeddings.

Adversarial Similarity Constraint: Inspired by the adversarial training (Chen et al., 2018; Duan, Zheng, Lin, Lu, & Zhou, 2018), we propose an adversarial similarity constraint to enhance the commonality between the modality-common representations from different modalities. Specifically, we treat the modality-common project function C as a generator and take the modality-common representations as generated results. Then we employ an MLP-based classifier as the discriminator D, which is used to distinguish the origin modality which the generated modality-common representation comes from. For the kth modality, let $C\left(\mathbf{e}_k\right)$ be the common representation from it. The probability that $C\left(\mathbf{e}_k\right)$ derives from the kth modality is as follows:

$$p_k = D\left(C\left(e_k\right)\right), \tag{4}$$

where p_k is the predicted value of the modality label. The modality discriminator tries to point out the original modality of the common representation, in other words, for $D(\cdot)$, to maximize p_k. However, the modality-common project function aims to generate consistent representations, in other words, for $C(\cdot)$, to minimize p_k. We optimize the generator and discriminator in an adversarial training manner until the discriminator has difficulty recognizing the difference between the extracted modality-common representations, indicating that the common representations from different modalities are sufficiently similar. The loss function is shown below:

$$\mathcal{L}_{\text{adv}} = \min_{\Theta^C} \max_{\Theta^D} \left(\ell_k \cdot \log D\left(C\left(\mathbf{e}_k\right)\right)\right), k \in \{t, v\}, \tag{5}$$

where ℓ_k denotes the ground-truth value of the modality label.

4.3.2. Purifying reconstruction decoder

Although the adversarial similarity constraints and the difference constraints ensure that the preliminary disentanglement encoder produces different representations, these simple constraints do not guarantee that the obtained representations are pure. The subsequent experimental section will illustrate that the modality-common and modality-specific subspaces produced by the preliminary disentanglement encoder are intermixed, which leads to the inability to obtain pure disentanglement representations. To address this issue, we design a purifying reconstruction decoder containing subtasks of intra-modal reconstruction and inter-modal reconstruction to obtain pure modality disentanglement representations. Specifically, the intra-modal reconstruction task aims to reconstruct the modality semantics utilizing common and specific representations within each modality. The inter-modal reconstruction task mutually reconstructs text and image semantics in a cross-modal manner. These two subtasks ensure that the obtained modality-common and modality-specific representations are pure and contain the full semantics of the news.

For intra-modal reconstruction, we feed the common and specific representations from the image(text) into decoders to reconstruct the original visual(textual) semantics as:

$$\bar{\mathbf{e}}_t^{\text{intra}} = \text{Dec}_T\left(\mathbf{h}_t^c + \mathbf{h}_t^s\right), \bar{\mathbf{e}}_v^{\text{intra}} = \text{Dec}_V\left(\mathbf{h}_v^c + \mathbf{h}_v^s\right), \tag{6}$$

where + denotes the vector addition operation, $\bar{\mathbf{e}}_t^{\text{intra}} \in \mathbb{R}^{d_c}$ and $\bar{\mathbf{e}}_v^{\text{intra}} \in \mathbb{R}^{d_c}$ have the same dimension as the original news representations $\mathbf{e}_t, \mathbf{e}_v$.

For inter-modal reconstruction, we reconstruct textual semantics using modality-specific representations from text and modality-common representations from images, and vice versa.

$$\bar{\mathbf{e}}_t^{\text{inter}} = \text{Dec}_T\left(\mathbf{h}_v^c + \mathbf{h}_t^s\right), \bar{\mathbf{e}}_v^{\text{inter}} = \text{Dec}_V\left(\mathbf{h}_t^c + \mathbf{h}_v^s\right). \tag{7}$$

The reconstruction loss is made up of the mean square error between the original representations and the decoder output representations, as follows:

$$\mathcal{L}_{\text{rec}} = \left\|\mathbf{e}_t - \bar{\mathbf{e}}_t^{\text{intra}}\right\|_2^2 + \left\|\mathbf{e}_t - \bar{\mathbf{e}}_t^{\text{inter}}\right\|_2^2 + \left\|\mathbf{e}_v - \bar{\mathbf{e}}_v^{\text{intra}}\right\|_2^2 + \left\|\mathbf{e}_v - \bar{\mathbf{e}}_v^{\text{inter}}\right\|_2^2, \tag{8}$$

where $\|\cdot\|_2^2$ represents the squared L_2−norm, and the total loss of the multimodal disentanglement autoencoder module is formulated as:

$$\mathcal{L}_{\text{DEC}} = \mathcal{L}_{\text{diff}} + \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{rec}}. \tag{9}$$

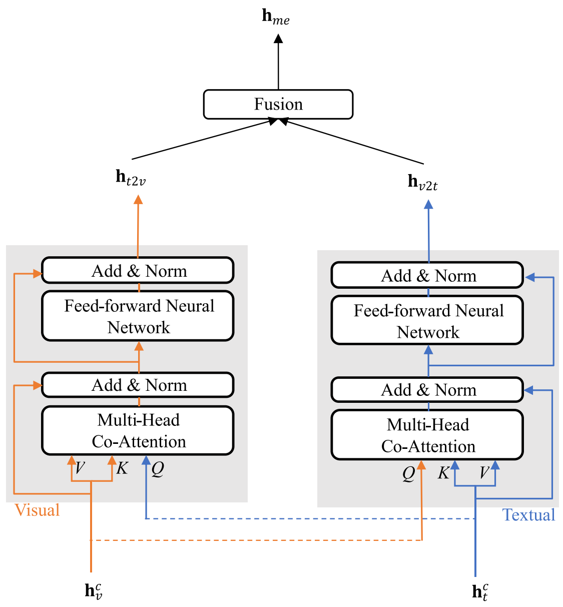


Fig. 3. Cross-modal mutual enhancement unit.

4.4. Cross-modal clues mining

After obtaining modality-common and modality-specific representations, the cross-modal clues mining module aims to mine multiple cross-modal clues explicitly and comprehensively.

4.4.1. Modality-common semantic enhancement

In multimodal fake news, modality-common semantics frequently contain important elements of the news event. Although common semantics do not provide additional information, different modalities can provide different views to enhance understanding of the news event. For example, the textual view provides context information about the news, while the visual view provides color and pattern information. Fusing different views can achieve enhanced semantic representations. Therefore, we propose a cross-modal mutual enhancement unit to model modality-common semantic enhancement.

Benefiting from the outstanding performance of the co-attention mechanism (Yang et al., 2019) in multimodal tasks, we use a co-attention transformer layer to fuse modality-common representations. As shown in Fig. 3, we feed the modality-common representations $\mathbf{h}_t^c$ and $\mathbf{h}_v^c$ into a two-stream transformer to process text and image information concurrently. Two streams output the language-enhanced image representation $\mathbf{h}_{t2v}$ and the visual-enhanced text representation $\mathbf{h}_{v2t}$, respectively. Each stream is equipped with a multi-headed co-attention block and a feed-forward neural network, both of which are followed by a residual connection and layer normalization. Notably, the dual streams share parameters to prevent overfitting.

Specifically, we first define a series of weight matrices $\mathbf{W}_j^{\mathcal{Q}}, \mathbf{W}_j^{\mathcal{K}}, \mathbf{W}_j^{\mathcal{V}} \in \mathbb{R}^{d \times d_m}$ to project the modality-common representations into the query, key, and value vectors, respectively:

$$\begin{aligned} &\mathbf{Q}_{v_j} = \mathbf{h}_v^c \mathbf{W}_j^{\mathcal{Q}}, \quad \mathbf{K}_{v_j} = \mathbf{h}_v^c \mathbf{W}_j^{\mathcal{K}}, \quad \mathbf{V}_{v_j} = \mathbf{h}_v^c \mathbf{W}_j^{\mathcal{V}}, \\ &\mathbf{Q}_{t_j} = \mathbf{h}_t^c \mathbf{W}_j^{\mathcal{Q}}, \quad \mathbf{K}_{t_j} = \mathbf{h}_t^c \mathbf{W}_j^{\mathcal{K}}, \quad \mathbf{V}_{t_h} = \mathbf{h}_t^c \mathbf{W}_j^{\mathcal{V}}, \end{aligned} \tag{10}$$

where $d_m = d/M$ is the dimension of one-head co-attention block, j denotes the jth head, and M is the number of heads.

Subsequently, we input query, key, and value vectors into a dual-stream multi-head co-attention block in order to compute the image-to-text and text-to-image attention vectors, which enables to model the cross-modal correlations. The co-attention block (Yang et al., 2019) is a modified version of the self-attention block found in the transformer (Vaswani et al., 2017). In this variant, the keys and values from one modality are utilized as inputs for the multi-headed attention block of the other modality to facilitate interactions between the modalities. The dot product between the query and key vectors determines the attention distribution on the value vector. To comprehensively investigate the intricate relationship between the image and the text from various perspectives, the multi-head attention mechanism employs distinct weight matrices to project the query, key, and value vectors for M times and subsequently process them concurrently. Finally, the processed results are concatenated and projected to obtain the representation of multi-head attention. The computation procedure for the multi-head co-attention function can be illustrated as follows:

$$\begin{cases} \mathrm{Att}_{t2v}^{j}\left(\mathbf{h}_t^c, \mathbf{h}_v^c\right) = \mathrm{softmax}\left(\dfrac{\mathbf{Q}_{t_j}\mathbf{K}_{v_j}^{\top}}{\sqrt{d_m}}\right)\mathbf{V}_{v_j}, \\ \text{MH-Att}_{t2v} = \left[\mathrm{Att}_{t2v}^{1}; \mathrm{Att}_{t2v}^{2} \dots ; \mathrm{Att}_{t2v}^{M}\right]\mathbf{W}', \end{cases} \tag{11}$$

$$\begin{cases} \mathrm{Att}_{v2t}^{j}\left(\mathbf{h}_v^c, \mathbf{h}_t^c\right) = \mathrm{softmax}\left(\dfrac{\mathbf{Q}_{v_j}\mathbf{K}_{t_j}^{\top}}{\sqrt{d_m}}\right)\mathbf{V}_{t_j}, \\ \text{MH-Att}_{v2t} = \left[\mathrm{Att}_{v2t}^{1}; \mathrm{Att}_{v2t}^{2} \dots ; \mathrm{Att}_{v2t}^{M}\right]\mathbf{W}', \end{cases} \tag{12}$$

where Att^j refers to the jth head of multi-head co-attention, $\mathbf{W}' \in \mathbb{R}^{d\times d}$ is the learnable weight matrix, and $[\cdot;\cdot]$ denotes the concatenation operation.

To enhance the model representation and training stability, we introduce the feed-forward blocks (FFN) with residual connection and layer normalization (LN) into the co-attention transformer layer. We obtain the visual-enhanced text representation $\mathbf{h}_{v2t}$ and language-enhanced image representation $\mathbf{h}_{t2v}$ through a two-stream co-attention transformer layer as below:

$$\begin{cases} \mathbf{h}_{v2t}' = \mathrm{LN}\left(\text{MH-Att}_{v2t} + \mathbf{h}_t^c\right), \\ \mathbf{h}_{v2t} = \mathrm{LN}\left(\mathrm{FFN}\left(\mathbf{h}_{v2t}'\right) + \mathbf{h}_{v2t}'\right). \end{cases} \tag{13}$$

$$\begin{cases} \mathbf{h}_{t2v}' = \mathrm{LN}\left(\text{MH-Att}_{t2v} + \mathbf{h}_v^c\right), \\ \mathbf{h}_{t2v} = \mathrm{LN}\left(\mathrm{FFN}\left(\mathbf{h}_{t2v}'\right) + \mathbf{h}_{t2v}'\right). \end{cases} \tag{14}$$

While some existing studies (Qian et al., 2021; Song et al., 2021; Wu et al., 2021) also use the co-attention mechanism to fuse multimodal information, their and our approaches are different. Specifically, they use coupled modality representations as inputs to the co-attention block. However, the modality-specific information inevitably introduces noise for cross-modal alignment. In contrast, we use refined modality-common representations as inputs, which avoids noise.

Finally, we fuse the language-enhanced image representation $\mathbf{h}_{t2v}$ and the visual-enhanced text representation $\mathbf{h}_{v2t}$ to obtain the multimodal mutual enhancement representation $\mathbf{h}_{me}$, as below:

$$\mathbf{h}_{me} = f_{\mathrm{fuse}}\left(\mathbf{h}_{v2t}, \mathbf{h}_{t2v}\right) = W_{\mathrm{fuse}}\left[\mathbf{h}_{v2t}; \mathbf{h}_{t2v}\right], \tag{15}$$

where $f_{\mathrm{fuse}}()$ is the fusion function, and $W_{\mathrm{fuse}} \in \mathbb{R}^{d\times 2d}$ is a transformation matrix.

4.4.2. Modality-specific semantic complementation

Multimodal data are complementary because each modality has its modality-specific semantics. The unique semantics from different modalities complement each other and thus enrich the information of the news. Unlike previous studies modeling complementation by directly concatenating features from different modalities that cause substantial information redundancy, we concatenate refined modality-specific representations as $\left[\mathbf{h}_t^s, \mathbf{h}_v^s\right]$ to model the multimodal complementary clue.

4.4.3. Modality-specific semantic inconsistency

Utilizing irrelevant images is a common strategy to fabricate fake news. Consequently, measuring inconsistency between modalities is a crucial way to identify fake news. We expect our cross-modal inconsistency comparison unit to focus more on the differences between modalities rather than on the commonalities. Therefore, we ignore modality-common representations and focus only on modality-specific representations. Specifically, we calculate the inconsistency comparison representation $\mathbf{h}_{mi}$ as below:

$$\mathbf{h}_{mi} = f_{\mathrm{cmp}}\left(\mathbf{h}_t^s, \mathbf{h}_v^s\right) = W_{\mathrm{cmp}}\left[\mathbf{h}_t^s - \mathbf{h}_v^s; \mathbf{h}_t^s \odot \mathbf{h}_v^s\right], \tag{16}$$

where $f_{\mathrm{cmp}}()$ represents the comparison function, $W_{\mathrm{cmp}} \in \mathbb{R}^{d\times 2d}$ is a transformation matrix, $-$ indicates the element-wise subtraction for computing semantic differences between modality-specific representations, and $\odot$ is Hadamard product, i.e., the element-wise product for computing semantics relationships between modality-specific representations.

In summary, the cross-modal clues mining module explores three cross-modal clues comprehensively and effectively, including modality-specific semantic complementation clue representation $\left[\mathbf{h}_t^s, \mathbf{h}_v^s\right]$, modality-common semantic enhancement clue representation $\mathbf{h}_{me}$ and modality-specific semantic inconsistency clue representation $\mathbf{h}_{mi}$. Furthermore, the next section will reweigh and aggregate these representations to detect fake news effectively and explainably.

4.5. Adaptive attention aggregation

Since various cross-modal clues have different effects on the prediction results, aggregating diverse representations based on their contributions can enhance detection and provide explainability. We propose an adaptive attention aggregation module to capture the contribution weights of different representations $\mathbf{h}_t^s, \mathbf{h}_v^s, \mathbf{h}_{me}$ and $\mathbf{h}_{mi}$ before their aggregation. More formally, we use one shared

Table 1
Statistics of three multimodal fake news detection datasets.

	Weibo	PHEME	Gossipcop
Fake News	4749	1972	1650
Real News	4779	3830	16 767

attention vector $\mathbf{q} \in \mathbb{R}^{d\times 1}$ to get the attention values μ_{ta} as follows:

$$\begin{cases} \mu_t = \mathbf{q}^\top \cdot \tanh\left(\mathbf{W}_{\text{att}}\left(\mathbf{h}_t^s\right) + \mathbf{b}_{\text{att}}\right), \\ \mu_v = \mathbf{q}^\top \cdot \tanh\left(\mathbf{W}_{\text{att}}\left(\mathbf{h}_v^s\right) + \mathbf{b}_{\text{att}}\right), \\ \mu_{me} = \mathbf{q}^\top \cdot \tanh\left(\mathbf{W}_{\text{att}}\left(\mathbf{h}_{me}\right) + \mathbf{b}_{\text{att}}\right), \\ \mu_{mi} = \mathbf{q}^\top \cdot \tanh\left(\mathbf{W}_{\text{att}}\left(\mathbf{h}_{mi}\right) + \mathbf{b}_{\text{att}}\right), \end{cases} \tag{17}$$

where $\mathbf{W}_{\text{att}} \in \mathbb{R}^{d\times d}$ and $\mathbf{b}_{\text{att}} \in \mathbb{R}^{d\times 1}$. Then, with a softmax-based normalization, we obtain the contribution weights $w = \{w_v, w_t, w_{me}, w_{mi}\}$ as below:

$$w_{ta} = \frac{\exp\left(\mu_{ta}\right)}{\sum_{ta\in\{t,v,mi,me\}} \exp\left(\mu_{ta}\right)}. \tag{18}$$

A large weight means the corresponding representation is crucial. We multiply the contribution weights with $\mathbf{h}_v$, $\mathbf{h}_t$, $\mathbf{h}_{me}$, and $\mathbf{h}_{mi}$ and sum them to obtain the final news representation $\mathbf{x} \in \mathbb{R}^d$ as

$$\mathbf{x} = w_v\mathbf{h}_v^s + w_t\mathbf{h}_t^s + w_{me}\mathbf{h}_{me} + w_{mi}\mathbf{h}_{mi}. \tag{19}$$

4.6. Classification

Eventually, we input the aggregated representation $\mathbf{x}$ into a classification layer to predict the news's truthfulness label $\hat{y}$.

$$\hat{y} = \text{softmax}\left(\mathbf{W}_{\text{cls}}\mathbf{x} + \mathbf{b}_{\text{cls}}\right), \tag{20}$$

where $\mathbf{W}_{\text{cls}}$ and $\mathbf{b}_{\text{cls}}$ are the learnable parameters. For the FND task, we optimize the parameters using the traditional cross-entropy (CE) loss function as below:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log \hat{y}_i, \tag{21}$$

where N represents the total count of news items, y_i and $\hat{y}_i$ refer to the ground-truth label and the predicted label, respectively. Combining the multimodal disentanglement autoencoder loss and the CE loss, the final training objectives are as below:

$$\mathcal{L}_{\text{ALL}} = \mathcal{L}_{\text{CE}} + \lambda\mathcal{L}_{\text{DEC}}, \tag{22}$$

where λ is a hyper-parameter to control the extent of disentanglement representation learning.

5. Experiment

In this section, we perform experiments on three public datasets to evaluate the performance and explainability of DCCMA-Net. In Section 5.1, we first provide an overview of the three datasets we utilize, then describe the comparative methods, and finally provide the implementation detail. Subsequently, in Section 5.2, we compare DCCMA-Net with other comparative methods on three public multimodal fake news detection datasets to evaluate its performance. In Section 5.3, we conduct quantitative experiments to evaluate the explainability of DCCMA-Net. In Section 5.4, we fully discuss the effectiveness of each innovation strategy of our model. In Section 5.5, we examine the sensitivity of DCCMA-Net to the hyperparameter. In Section 5.6, we further visualize the effectiveness and explainability of DCCMA-Net on the fake news detection task by analyzing several typical cases.

5.1. Experimental setup

5.1.1. Dataset

To evaluate the performance and explainability of the DCCMA-Net model, we conduct experiments on three public datasets, namely Weibo (Jin et al., 2017), PHEME (Zubiaga et al., 2017), and Gossipcop (Shu et al., 2020). The statistical information of the datasets is shown in Table 1. The details of the datasets are described below:

- **Weibo** dataset is proposed by Jin et al. (2017) and is extensively utilized in the Chinese multimodal fake news detection task. The Weibo dataset comprises Chinese news articles along with their corresponding images. The dataset is labeled as true and false by professional editors and experts, where true news is sourced from Xinhua News Agency[3] while fake news is crawled from Weibo[4] platform. Following the benchmark (Jin et al., 2017) setting, we divided the dataset into training and test sets in the ratio of 8:2.
- **PHEME** is a widely used English multimodal fake news detection dataset, which is proposed by Zubiaga et al. (2017). The dataset is collected based on five breaking events from the Twitter[5] platform. The PHEME dataset contains Twitter text and the corresponding images. The news items in the dataset are labeled as true and false by senior editors and experts. Following previous work (Qian et al., 2021), the PHEME dataset is divided into training and testing sets with a ratio of 8:2.
- **Gossipcop** dataset is proposed by Shu et al. (2020), which is collected at the famous fact-checking website Gossipcop.[6] Multimodal news articles in the Gossipcop dataset are published from July 2000 to December 2018. Domain experts label the news in the dataset as true and false. Following previous work (Wei et al., 2022; Zhou et al., 2020), we divide the Gossipcop dataset into training and testing sets in the ratio of 8:2.

5.1.2. *Comparative methods*

In order to verify the performance of DCCMA-Net, we choose the state-of-the-art automatic fake news detection methods as the comparative methods, which include text-based methods and multimodal methods. We do not perform manual fake news detection experiments. People with different knowledge levels have different capabilities to detect false news, so it is difficult to verify the average level of manual detection. Notably, the datasets we selected are labeled by domain experts, which can represent the ability of human experts to detect fake news. In this subsection, we detail the comparative methods as below.

Text-based methods.

- **DTC** (Castillo et al., 2011) feeds the manually extracted text features into the decision tree model to obtain the classification results.
- **GRU** (Ma et al., 2016) utilizes GRU to learn news text features for classification.
- **BERT** (Kenton & Toutanova, 2019) utilizes a pre-trained BERT model to learn news text features for classification. Specifically, we use *chinese-bert-wwm-ext* and *bert-base-uncased* from Transformers package (Wolf et al., 2020) for the Chinese and English evaluation, respectively. And we test BERT after fine-tuning it on the fake news detection dataset.
- **GPT-3.5** (OpenAI, 2023) is a large language model proposed by open AI. GPT-3.5 performs well in natural language understanding and natural language generation. Hu et al. (2024) design the prompt to stimulate the potential of GPT-3.5 for fake news detection.
- **GPT-4** (Achiam et al., 2023) is an upgraded version of GPT-3.5.

Multimodal methods.

- Joint Representation Method: **SpotFake** (Singhal et al., 2019) uses pre-trained BERT (Kenton & Toutanova, 2019) and VGG19 (Simonyan & Zisserman, 2015) models to extract text and image features, respectively, and concatenates them for classification. **SpotFake**+ (Singhal et al., 2020) extracts text and image features using pre-trained XLNet (Yang et al., 2019) and VGG19, respectively, and concatenates them for classification. **EANN** (Wang et al., 2018) concatenates text and image features as multimodal news representations. Furthermore, EANN introduces an event classification auxiliary task to help the model obtain better news representations. **MVAE** (Khattar et al., 2019) concatenates text and image features as multimodal news representations. MVAE introduces a modality reconstruction auxiliary task to learn better multimodal representations. **CMC** (Wei et al., 2022) presents a knowledge distillation module to enhance the unimodal encoder. After that, CMC fed the unimodal representations into the bi-linear pooling module to obtain the multimodal news representations for classification. **LIIMR** (Singhal et al., 2022) proposes a multiplicative multimodal method to capture the contribution weights of different modalities. LIIMR aggregates the features of image and text based on the contribution weights. LIIMR outputs contribution weights as explanations for the detection results.
- Alignment Representation Method: **att_RNN** (Jin et al., 2017) utilizes an attention-based RNN network to align text tokens with the image. **MCAN** (Wu et al., 2021) employs a co-attention network to align features from different modalities. **CARMN** (Song et al., 2021) proposes a cross-modal attention residual network to fuse multimodal features and design a multi-channel CNN to mitigate the noise generated during the modal fusion process. **HMCAN** (Qian et al., 2021) employs a hierarchical attention model that considers hierarchical text and cross-modal contextual semantics. **BTIC** (Zhang, Gui et al., 2021) employs a self-attention mechanism to align multimodal features. In addition, BTIC utilizes supervised contrastive loss and cross-entropy loss to jointly optimize the model. **MPFN** (Jing et al., 2023) proposes a progressive fusion network to align deep and shallow features from two modalities simultaneously. **MMFN** (Zhou et al., 2023a) fuses multimodal multi-granularity information for fake news detection.

[3] http://www.xinhuanet.com.

[4] https://weibo.com.

[5] https://twitter.com/.

[6] https://wwwgossipcop.com/.

- Comparison Representation Method: **SAFE** (Zhou et al., 2020) utilizes an image caption model to convert the image to text and measures cross-modal inconsistency by comparing the similarity of the original news text with the generated text. **MCNN** (Xue et al., 2021) maps different modality features to the same embedding space using a parameter-sharing projection network. Then MCNN compares the inconsistencies between different modalities. **CAFE** (Chen et al., 2022) measures inconsistency by computing the KL divergence between the representations of different modalities. **DGExplain** (Shang et al., 2022) utilizes a dual-generative cross-modal feature generation approach to concurrently produce textual-guided visual and image-guided textual features. Then, DGExplain measures cross-modal inconsistency by comparing the generated features with the original news features. DGExplain predicts news veracity using each content feature (including image, text, and cross-modal inconsistency) and uses the prediction score as an explanation for the detection result. **FND-CLIP** (Zhou, Yang, Ying, Qian, & Zhang, 2023b) utilizes the CLIP similarity score to measure cross-modal inconsistency. The CLIP score also guides the model to use different modality data. **COOLANT** (Wang et al., 2023) proposes a cross-modal contrastive learning mechanism to project the image and the text into the unified embedding space. Then, the representations are used to measure the image–text inconsistency. Finally, COOLANT proposes a cross-modal aggregation mechanism to reweight and aggregate different content features (including image, text, and cross-modal inconsistency). COOLANT outputs aggregation weights as explanations for the detection result.
- Hybrid Method: **EM-FEND** (Qi et al., 2021) extracts text and visual entities and mines entity-level enhancement and inconsistency associations. At the same time, EM-FEND extracts image-embedded text as a complement to the news text. **BMR** (Ying et al., 2023) extracts news features from views of text, image pattern, image semantics, and modality consistency and proposes an improved multi-gate mixture-of-expert network to reweight and aggregate them adaptively.

5.1.3. *Implementation details*

The experiments were conducted on a server equipped with four GeForce RTX 3090Ti GPUs. We utilize accuracy and F1 score, as our evaluation metrics on three datasets. To ensure a fair comparison, we compare the performance of our proposed DCCMA-Net with other comparative methods under the same dataset and the same data preprocessing conditions. Furthermore, to mitigate the impact of errors on experimental outcomes, we ran each experiment ten times and take the mean value as the final result.

For our proposed DCCMA-Net model, we use the pre-trained CLIP (Radford et al., 2021) model to extract text and image features. Due to the lack of a pre-trained Chinese model in CLIP, we employ the Google Translation API[7] to translate Chinese text into English. To comply with the input size constraints of CLIP, we utilize a summary generation model (Raffel et al., 2020) to produce concise summaries for texts that surpass 50 words. The CLIP model we utilize is “ViT-B/32” with a feature dimension of $d_c = 512$. To prevent overfitting, we freeze the parameters of CLIP during the training process. In the multimodal disentanglement autoencoder module, we set the dimension of the modality-common and -specific representations as $d = 256$. In the cross-modal clues mining module, we set the co-attention layer with eight attention heads. In the training phase, we set the disentanglement coefficient λ as 0.3, the mini-batch size as 128, and the learning rate as $5e-4$. The training epoch is set to 100, and an early termination mechanism is included to prevent overfitting. Meanwhile, we utilize the Adam algorithm (Kingma & Ba, 2015) to optimize the parameters. The code for DCCMA-Net will be released on https://github.com/Snowsiqi/DCCMA-Net.

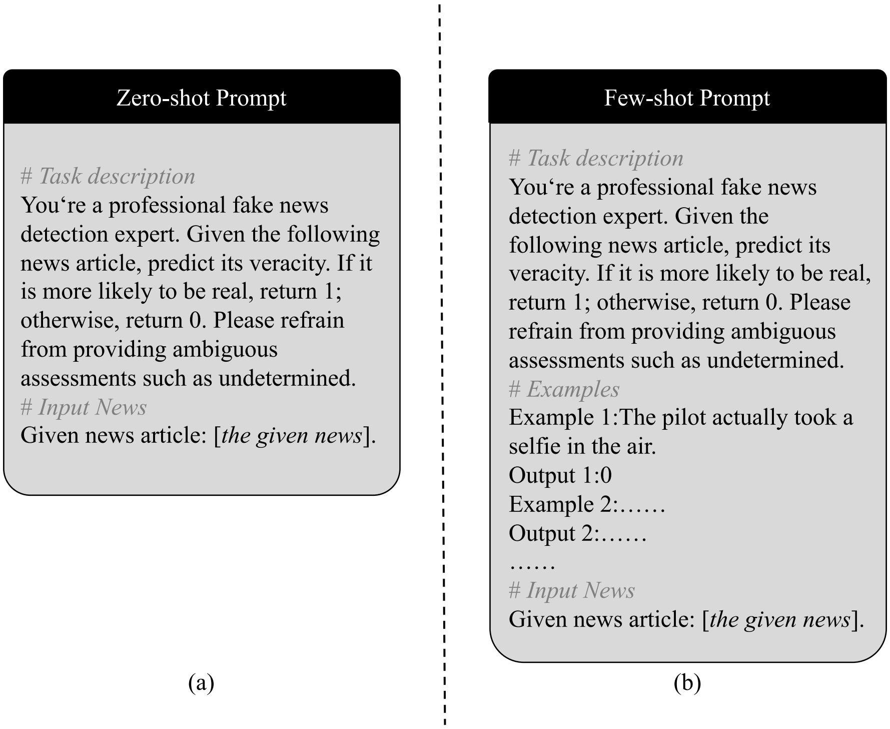
For comparative methods, we all use the parameter settings in the original paper to ensure the best performance. For EM-FEND, we intercepted the experimental results on the Weibo dataset from the original paper. This is because EM-FEND uses the Entity Extraction API to extract entities from news images, but the original paper does not publicize the API settings. We test the performance of large language models (e.g., GPT3.5 and GPT4) on the fake news detection task. Due to the large number of parameters and the unavailability of pre-training weights, task-specific fine-tuning of GPT3.5 and GPT4 becomes challenging. Therefore, we employ prompt learning techniques to inspire the potential of LLMs in the fake news detection task. Specifically, LLMs learn the fake news detection task through prompts that contain instructions or a small number of instances. Specifically, we use the following two prompting approaches to inspire the potential of LLMs on the fake news detection task.

- Zero-shot prompting: Zero-shot prompts contain only the task description and the given news. To achieve more accurate detection, we use role-playing techniques (Wu et al., 2024) when describing the task. The specific zero-shot prompt is shown in Fig. 4(a).
- Few-shot prompting: Few-shot prompts contain the task description, several news-label examples and the given news. Limited by the length of the LLM input field, we randomly select four examples, two of which are real news and two of which are fake news. It should be noted that since the samples in the dataset are independently and identically distributed, randomly selecting the examples does not affect the performance. The specific few-shot prompt is shown in Fig. 4(b)

5.2. *Performance comparison with state-of-art methods*

We compare the performance of DCCMA-Net with other state-of-the-art fake news detection methods on each dataset. The results of the comparison are displayed in Table 2. We can draw the following conclusions.

[7] http://translate.Google.com.



Zero-shot Prompt

Task description
You‘re a professional fake news detection expert. Given the following news article, predict its veracity. If it is more likely to be real, return 1; otherwise, return 0. Please refrain from providing ambiguous assessments such as undetermined.
Input News
Given news article: [*the given news*].

(a)

Few-shot Prompt

Task description
You‘re a professional fake news detection expert. Given the following news article, predict its veracity. If it is more likely to be real, return 1; otherwise, return 0. Please refrain from providing ambiguous assessments such as undetermined.
Examples
Example 1:The pilot actually took a selfie in the air.
Output 1:0
Example 2:……
Output 2:……
……
Input News
Given news article: [*the given news*].

(b)

Fig. 4. Prompts for LLMs conducting fake news detection tasks.

- DCCMA-Net outperforms other comparative methods in terms of ACC and F1 metrics on all three datasets. Our proposed DCCMA-Net achieves an accuracy of 0.931, 0.933, and 0.913 on the Weibo, PHEME, and Gpssipcop datasets, respectively, which suggests that DCCMA-Net is close to the detection ability of human experts. This is attributed to the following three reasons: Firstly, DCCMA-Net decomposes each modality into modality-specific and modality-common representations, which effectively mitigates the redundancy. Second, DCCMA-Net comprehensively and explicitly mines three cross-modal clues to provide evidence for fake news detection. Finally, DCCMA-Net adaptively reweights and aggregates different clues to produce highly discriminative news representations.
- Text-based methods perform weaker than multimodal methods, which indicates that the introduction of multimodal information for fake news detection is necessary. Among the text-based methods, the DTC, which is based on hand-extracted features, performs weaker than the deep learning-based methods (GRU, BERT, GPT3.5, and GPT4), which suggests that hand-extracted features have a weak generalization performance. BERT outperforms GRU, which suggests that the contextual comprehension capability developed by the pre-training process is effective for understanding news content. However, despite pre-training with a larger corpus, LLMs (both GPT3.5 and GPT4) underperform the fine-tuned BERT using all two prompting approaches. We analyze the main reasons as follows: First, the fake news detection task requires the model to have a strong ability to detect fake features such as forgery and non-facts. Although LLMs have strong contextual understanding and natural language generation capabilities, LLMs are weak at capturing fake features due to the lack of task-specific adaptations. However, BERT enhances the ability to capture fake features by fine-tuning on the fake news detection dataset. Thus the fine-tuned BERT performs stronger than LLMs. Second, LLMs are limited by challenges in factuality and hallucinations, which also leads to weaker performance when using LLMs directly for fake news detection. In addition, we analyze the reasons that affect the fake news detection performance of LLMs: First, knowledge and capability of the LLMs itself: We found that GPT4 outperforms GPT3.5 when using different prompt methods. GPT, an updated version of GPT 3.5, has richer knowledge and better capabilities. Therefore, the knowledge and capability of LLMs itself is correlated with their performance on the fake news detection task. Second, design of prompts: we find that the few-shot prompting methods perform better than the zero-shot prompting methods. This is because few-shot methods introduce some task-specific examples that can help the models improve the ability of capturing fake features. Therefore, the design of prompts is critical for LLMs to improve the performance of downstream tasks.
- DCCMA-Net outperforms the joint representation approaches on all three datasets. This is because joint representation methods only consider images as a complement to text but ignore other rich cross-modal clues and redundancy. LIIMR aggregates the image and text features based on their contributions and achieves good results. This shows that different modalities have different effects on the detection results. Inspired by this, we design the adaptive attention aggregation mechanism to aggregate features of different modalities according to their contributions.

Table 2
Results of comparison between DCCMA-Net and other fake news detection methods on three datasets. We evaluate the performance of the model using ACC and F1. Bold numbers in the table indicate the best performance. Stars demonstrate that the p-value is below 0.05 using t-test compared with the best comparative method BMR.

	Method	Weibo		PHEME		Gossipcop	
		ACC	F1	ACC	F1	ACC	F1
Text-based method	DTC (Castillo et al., 2011)	0.532	0.521	0.541	0.533	0.562	0.659
	GRU (Ma et al., 2016)	0.678	0.675	0.689	0.693	0.692	0.689
	BERT (Kenton & Toutanova, 2019)	0.685	0.671	0.719	0.703	0.692	0.709
	GPT-3.5 (Zero-Shot) (OpenAI, 2023)	0.603	0.588	0.581	0.602	0.605	0.601
	GPT-3.5 (Few-Shot) (OpenAI, 2023)	0.612	0.593	0.599	0.611	0.615	0.609
	GPT-4 (Zero-Shot) (Achiam et al., 2023)	0.631	0.652	0.681	0.669	0.664	0.651
	GPT-4 (Few-Shot) (Achiam et al., 2023)	0.654	0.661	0.692	0.682	0.679	0.691
Joint representation method	EANN (Wang et al., 2018)	0.779	0.781	0.761	0.767	0.753	0.758
	MVAE (Khattar et al., 2019)	0.782	0.784	0.751	0.763	0.774	0.771
	SpotFake (Singhal et al., 2019)	0.794	0.791	0.783	0.781	0.781	0.783
	SpotFake+ (Singhal et al., 2020)	0.796	0.797	0.791	0.792	0.789	0.792
	CMC (Wei et al., 2022)	0.868	0.871	0.839	0.834	0.844	0.852
	LIIMR (Singhal et al., 2022)	0.858	0.863	0.816	0.818	0.851	0.841
Alignment representation method	att-RNN (Jin et al., 2017)	0.779	0.789	0.798	0.795	0.789	0.787
	MCAN (Wu et al., 2021)	0.868	0.865	0.858	0.857	0.847	0.841
	CARMN (Song et al., 2021)	0.865	0.846	0.862	0.859	0.851	0.849
	HMCAN (Qian et al., 2021)	0.876	0.874	0.869	0.871	0.858	0.842
	BTIC (Zhang, Gui et al., 2021)	0.884	0.883	0.855	0.861	0.863	0.869
	MPFN (Jing et al., 2023)	0.877	0.875	0.861	0.864	0.859	0.849
	MMFN (Zhou et al., 2023a)	0.886	0.891	0.891	0.895	0.852	0.856
Comparison representation method	SAFE (Zhou et al., 2020)	0.811	0.813	0.811	0.816	0.831	0.825
	MCNN (Xue et al., 2021)	0.823	0.816	0.801	0.812	0.821	0.809
	CAFE (Chen et al., 2022)	0.845	0.839	0.831	0.829	0.849	0.851
	DGExplain (Shang et al., 2022)	0.857	0.849	0.849	0.851	0.842	0.846
	FND-CLIP (Zhou et al., 2023b)	0.859	0.856	0.847	0.845	0.849	0.856
	COOLANT (Wang et al., 2023)	0.896	0.891	0.866	0.859	0.855	0.849
Hybrid method	EM-FEND (Qi et al., 2021)	0.904	0.901	–	–	–	–
	BMR (Ying et al., 2023)	0.908	0.903	0.897	0.898	0.878	0.875
Our Method	DCCMA-Net	**0.931***	**0.929***	**0.933***	**0.932***	**0.913***	**0.911***

- Although the alignment representation methods achieve some results, they still perform weaker than DCCMA-Net, because they overlook the noise caused by modality-specific semantics during the alignment process.
- DCCMA-Net performs better than the comparison representation methods on each dataset. This is because the comparison representation methods directly measure the inconsistency between the entire image and the complete text without considering the impact of modality-common semantics on the comparison process.
- Although hybrid methods achieve some results by considering multiple cross-modal clues, they are still weaker than DCCMA-Net. This is because they do not distinguish between modality-specific and modality-common semantics, but rather treat individual modalities as an entangled whole. Entangled representations not only cause redundancy but also inevitably introduce noise for cross-modal clues mining. In contrast, DCCMA-Net disentangles each modality into modality-common and modality-specific representations to mitigate redundancy. Based on these disentangled representations, DCCMA-Net mines three cross-modal clues explicitly and completely. In addition, DCCMA-Net adaptively aggregates these clues to obtain highly discriminative news representations. Therefore, DCCMA-Net can achieve the state-of-the-art performance.

5.3. *Explainability comparison with state-of-art methods*

We validate the explainability of DCCMA-Net by comparing DCCMA-Net with other baseline methods in terms of the accuracy of generating explanations for the detection results. We compare DCCMA-Net with LIIMR (Singhal et al., 2022), DGExplain (Shang et al., 2022), BMR (Ying et al., 2023), and COOLANT (Wang et al., 2023), which are the only baselines that can provide explainability by generating the attention weight of each cross-modal clue. In particular, we randomly selected 200 news articles from the test set of the Weibo dataset. For each news article, we obtain the explanation by retrieving the cross-modal clue with the highest attention weight from DCCMA-Net and each baseline method.

Due to the lack of interpretable data annotations in the original dataset, we label fake news at a more fine-grained level based on the explanation of why the news is fake. Specifically, we label the selected fake news in the Weibo test set as image-specific semantic fake news, text-specific semantic fake news, modality-common semantic fake news, and cross-modal inconsistency fake news. To ensure the accuracy of the labeling, we engage five senior experts in the field of journalism to perform the labeling task and select the highest value as the definitive label. In the event of a tie in the voting, we will request the participation of additional experts until a clear label is obtained. We take the manually annotated explanations as ground-truth labels for explainability experiments.

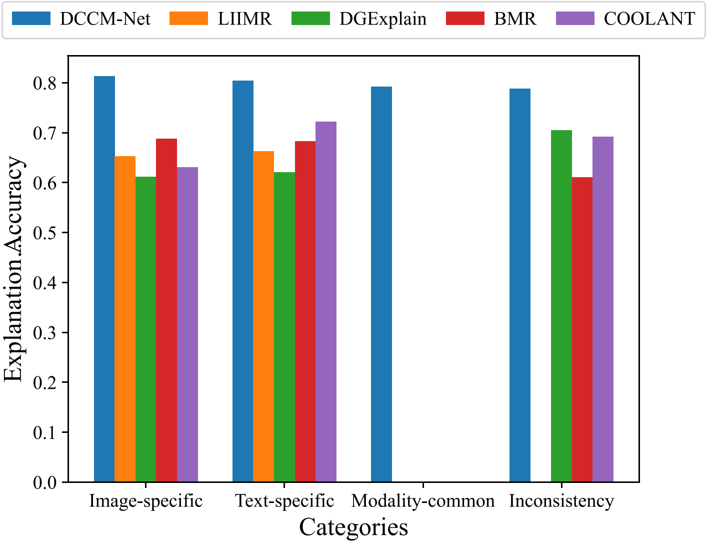


Fig. 5. Explainability performance.

After obtaining ground-truth explainability labels, we compare the accuracy of the explanations generated by DCCMA-Net and other baseline methods. We summarize the experiment results in Fig. 5. The experiment results show that DCCMA-Net outperforms other state-of-the-art interpretable methods in explaining the detection results. This is mainly attributed to the following two reasons. First, DCCMA-Net can explicitly explore multiple cross-modal clues to provide rich evidence for detecting fake news. Second, DCCMA-Net can reweight different clues and highlight the most contributive ones.

Meanwhile, we can see that other approaches cannot provide a comprehensive explanation for the different types of fake news. This is because other methods cannot comprehensively and explicitly mine cross-modal clues. In contrast, DCCMA-Net can comprehensively explore three cross-modal clues. Thus DCCMA-Net can explain multiple kinds of fake news.

5.4. Analysis

In this section, we perform detailed experiment analysis to understand our proposed DCCMA-Net in depth. First, we evaluate the impact of each key component on model performance. Then, we validate the efficacy of different constraints on the multimodal disentanglement autoencoder model. Finally, we evaluate the effectiveness of different cross-modal clues.

5.4.1. Effectiveness of each component

To verify the effectiveness of each component, we design the following variants of DCCMA-Net:

- w/o CLIP: In the modality feature extraction step, we replace the CLIP model with two unimodal encoders (Vit (Dosovitskiy et al., 2020) and BERT (Kenton & Toutanova, 2019)) to extract image and text features, respectively.
- w/o MDAE: We remove the multimodal disentanglement autoencoder module. Instead, we use the entangled unimodal representations to explore cross-modal clues.
- w/o CCM: We remove the cross-modal clues mining module. Instead, we concatenate modality-specific and modality-common representations for classification.
- w/o AAA: We remove the adaptive attention aggregation module. Instead, we regard the three cross-modal clues as equally important.

Quantitative analysis. We first compare the performance of DCCMA-Net and its variants on three datasets. The experiment results are presented in Table 3. The performance of DCCMA-Net drops after removing each component, proving that each component contributes to the model. When replacing CLIP encoders with unimodal pre-training models, the performance decreases. The reason is that CLIP aligns images and text in a uniform embedding space via contrastive learning, which helps mitigate modality heterogeneity. When the multimodal disentanglement autoencoder module is removed, the model cannot distinguish between modality-specific and modality-common semantics, which not only causes information redundancy but also brings noise to the subsequent cross-modal clues mining. So the performance decreases. After removing the cross-modal clues mining module, the model fails to mine the rich cross-modal clues, so the performance decreases. After removing the adaptive attention aggregation module, the model loses its ability to adaptively aggregate each representation based on its contribution. Therefore, the performance drops.

Table 3
Results of comparison between DCCMA-Net and its variant models on the three datasets. Bold numbers in the table indicate the best performance.

Method	Weibo		PHEME		Gossipcop	
	ACC	F1	ACC	F1	ACC	F1
DCCMA-Net	**0.931**	**0.929**	**0.933**	**0.932**	**0.913**	**0.911**
Effect of each component of the model						
DCCMA-Net w/o CLIP	0.926	0.923	0.928	0.929	0.909	0.907
DCCMA-Net w/o MDAE	0.919	0.915	0.916	0.914	0.901	0.902
DCCMA-Net w/o CCM	0.909	0.907	0.921	0.924	0.903	0.901
DCCMA-Net w/o AAA	0.921	0.919	0.922	0.923	0.905	0.904
Effect of each constraint						
DCCMA-Net w/o $\mathcal{L}_{\text{diff}}$	0.922	0.921	0.924	0.923	0.904	0.904
DCCMA-Net w/o $\mathcal{L}_{\text{adv}}$	0.921	0.917	0.925	0.924	0.903	0.902
DCCMA-Net w/o $\mathcal{L}_{\text{rec}}$	0.921	0.916	0.919	0.918	0.902	0.903
Effect of different cross-modal clues						
DCCMA-Net w/o $\mathbf{h}_{me}$	0.921	0.917	0.921	0.917	0.904	0.903
DCCMA-Net w/o $\mathbf{h}_{mi}$	0.919	0.922	0.924	0.924	0.902	0.906
DCCMA-Net w/o $\mathbf{h}_t^s + \mathbf{h}_v^s$	0.918	0.916	0.919	0.918	0.901	0.901

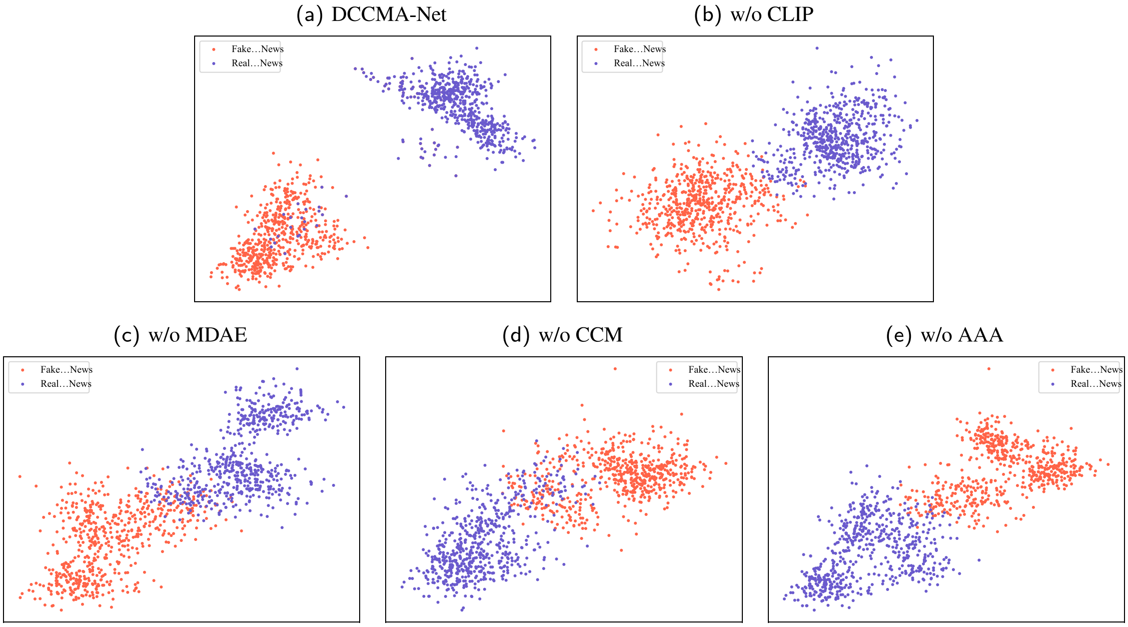


Fig. 6. T-SNE visualizations of the features before classifier that are learned by (a) DCCMA-Net, (b) DCCMA-Net w/o CLIP, (c) DCCMA-Net w/o MDAE, (d) DCCMA-Net w/o CCM, and (e) DCCMA-Net w/o AAA on the test dataset of Weibo. Dots with the same color are within the same label. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Visualization analyze. To analyze the impact of each component more intuitively, We perform T-SNE visualization (Van der Maaten & Hinton, 2008) of the features learned by DCCMA-Net and its variants. Specifically, we perform news feature visualization experiments on all three datasets. Figs. 6–8 show the results of T-SNE visualization of news features learned from DCCMA-Net and its variants on the Weibo, PHEME and Gossipcop datasets, respectively.

We compare the T-SNE visualization results for DCCMA-Net and its four variants. We can see that the classification boundary of different labeled points in DCCMA-Net is more obvious than its variants, indicating that each component in DCCMA-Net contributes to the model. Specifically, the features learned by DCCMA-Net w/o MDAE are less easy to categorize compared to the features learned by DCCMA-Net. The reason is that the MDAE module decomposes each modality into modality-common and -specific representations, which mitigates information redundancy and aids in cross-modal clues mining. After removing the cross-modal clues mining module, we can find that the classification boundary between real and fake news is more blurred, which demonstrates that three cross-modal clues can provide evidence for fake news detection. In addition, after removing the AAA layer, we can find that the learned features are not easily distinguishable. This suggests that adaptively aggregating multiple representations can significantly enhance the detection capacity of the model.

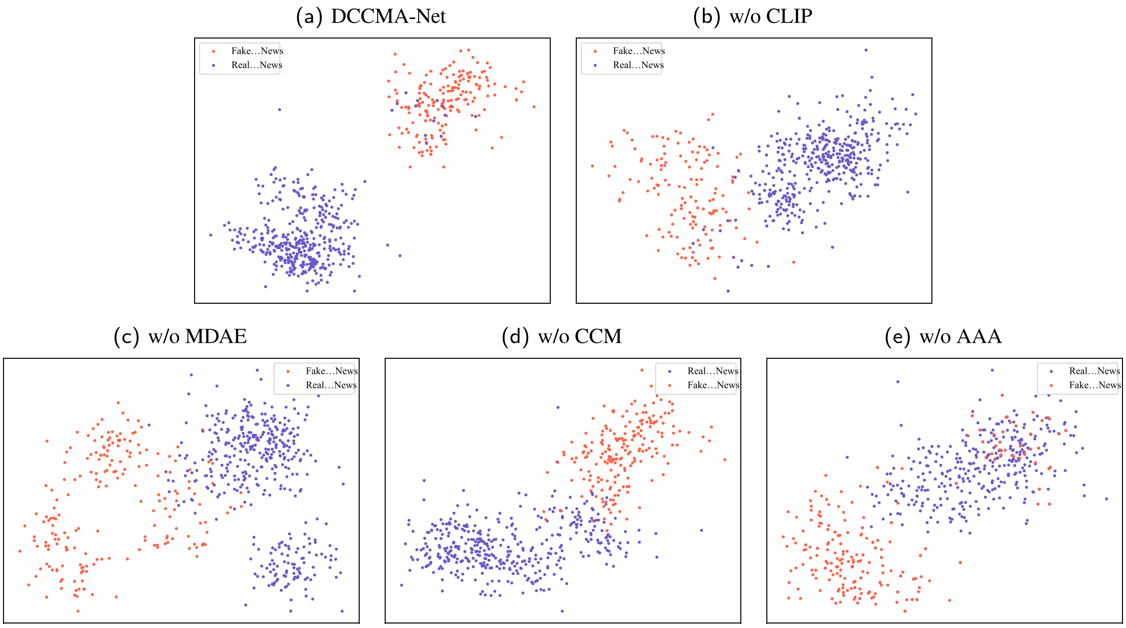


Fig. 7. T-SNE visualizations of the features before classifier that are learned by (a) DCCMA-Net, (b) DCCMA-Net w/o CLIP, (c) DCCMA-Net w/o MDAE, (d) DCCMA-Net w/o CCM, and (e) DCCMA-Net w/o AAA on the test dataset of PHEME. Dots with the same color are within the same label. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

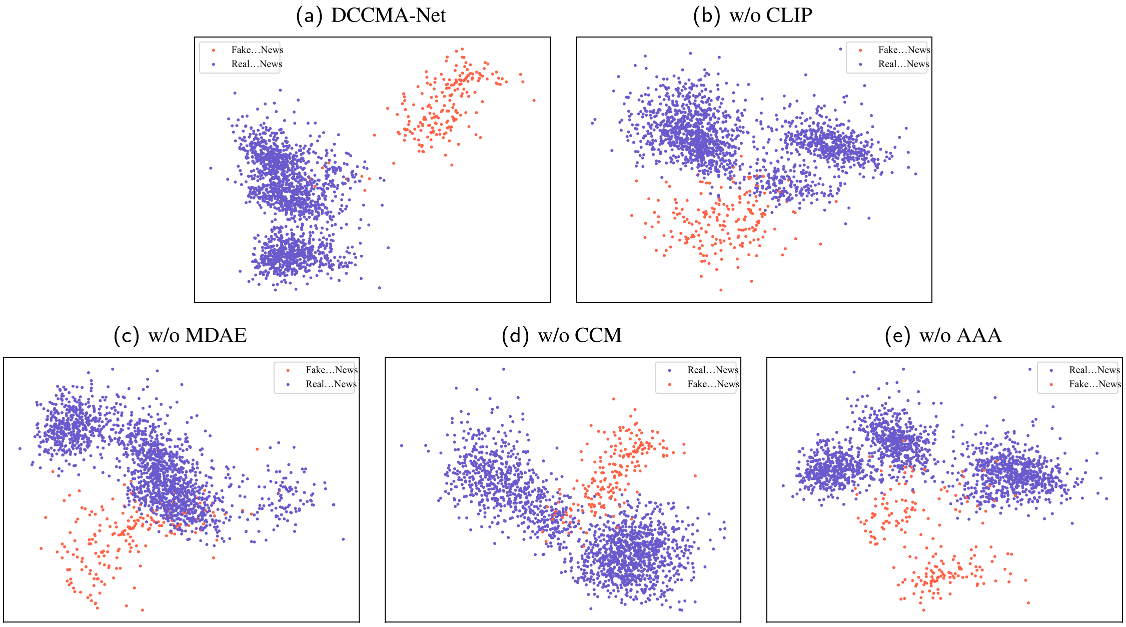
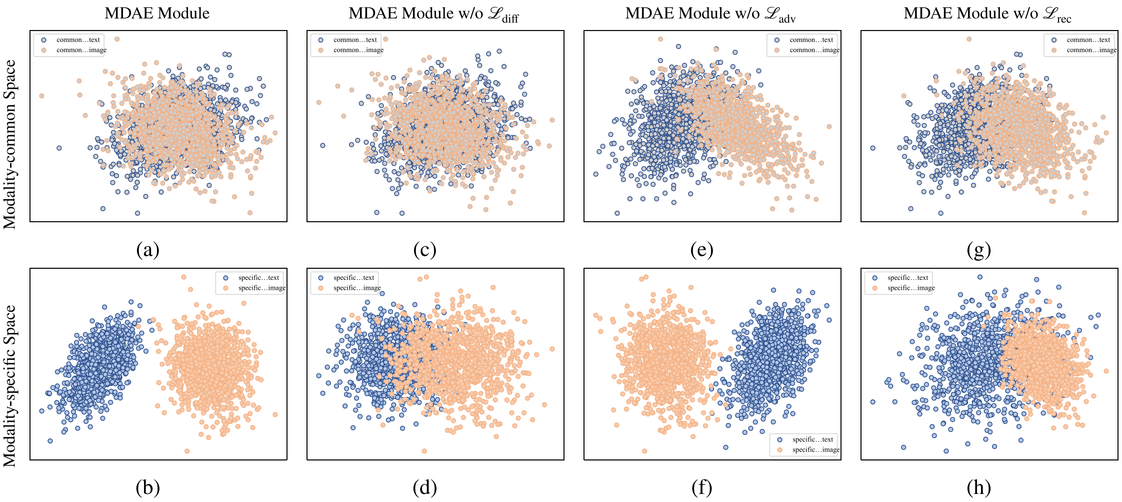


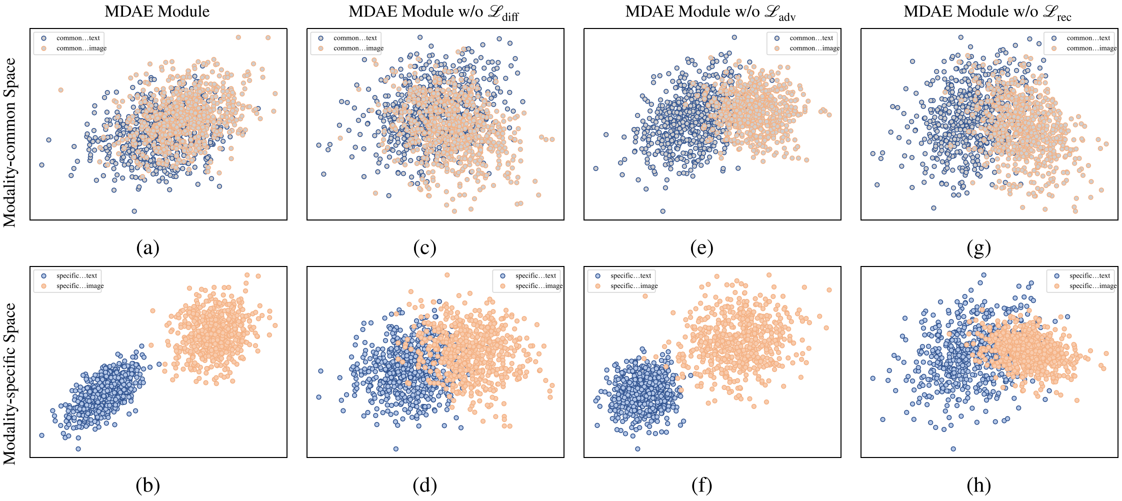
Fig. 8. T-SNE visualizations of the features before classifier that are learned by (a) DCCMA-Net, (b) DCCMA-Net w/o CLIP, (c) DCCMA-Net w/o MDAE, (d) DCCMA-Net w/o CCM, and (e) DCCMA-Net w/o AAA on the test dataset of Gossipcop. Dots with the same color are within the same label. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) (c) (e) (g)

(b) (d) (f) (h)

Fig. 9. Visualization of the modality-common and-specific subspaces learned by DCCMA-Net and its variants on the testing set of Weibo using T-SNE projection.



(a) (c) (e) (g)

(b) (d) (f) (h)

Fig. 10. Visualization of the modality-common and-specific subspaces learned by DCCMA-Net and its variants on the testing set of PHEME using T-SNE projection.

5.4.2. Effectiveness of each constraint

To verify the contribution of different constraints to the multimodal disentanglement autoencoder module, we design the following variants:

- w/o $\mathscr{L}_{diff}$: we remove the difference constraint.
- w/o $\mathscr{L}_{adv}$: we remove the adversarial similarity constraint.
- w/o $\mathscr{L}_{rec}$: we remove the reconstruction constraint.

Quantitative analysis. At first, we compare the performance of DCCMA-Net and its variants on three datasets. The comparison results are presented in Table 3. We can observe that the model performance decreases after removing any of the constraints. The three constraints together ensure deep disentanglement. Specifically, the difference constraint enables the separation of modality-specific semantics from modality-common semantics. The adversarial similarity constraints ensure the commonality of modality-common semantics. The reconstruction constraint guarantees the completeness and purity of the disentangled representations. Consequently, removing any of the constraints affects the disentanglement. Entangled representations result in redundancy and introduce noise for cross-modal clues mining. As a result, performance decreases.

Visualization analyze. To intuitively demonstrate the impact of various constraints on the MDAE module, we further visualize the modality-common and -specific representations acquired by the MDAE module and its variants on three test datasets. Figs. 9–11 show the results of T-SNE visualization of modality-common and modality-specific subspaces learned from the MDAE module and

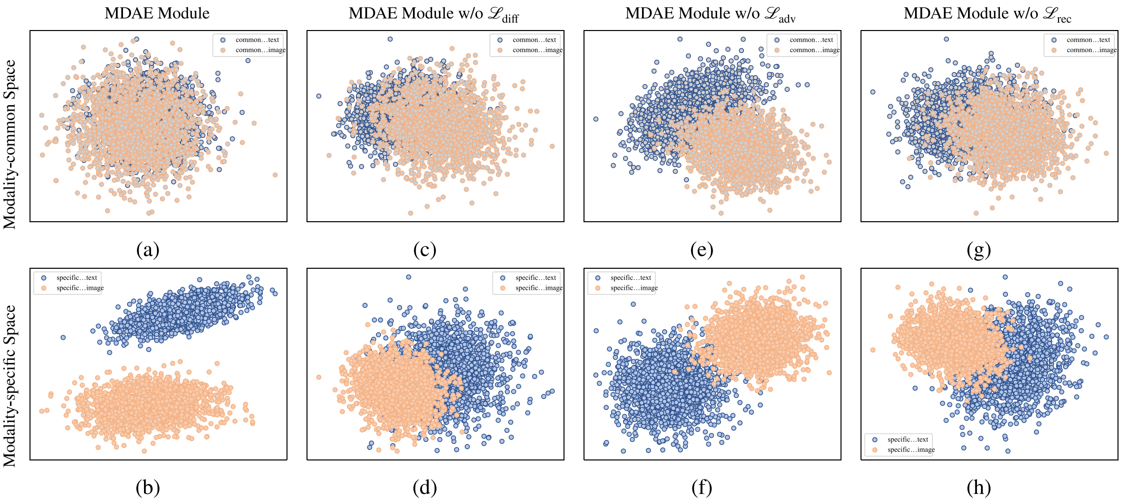


Fig. 11. Visualization of the modality-common and-specific subspaces learned by DCCMA-Net and its variants on the testing set of Gossipcop using T-SNE projection.

its variants on the Weibo, PHEME and Gossipcop datasets, respectively. The following conclusions can be inferred:

- The distributions of the modality-common subspace learned by the MDAE module are blended together. Moreover, the boundary of the modality-specific subspace learned by the MDAE module is obvious. And the distributions of modality-specific subspaces are more compact. This indicates that the MDAE module can learn the disentanglement representations well.
- After removing $\mathscr{L}_{\text{diff}}$, the text-specific and image-specific subspaces overlap. This suggests that the model does not learn modality-specific semantics well.
- After removing $\mathscr{L}_{\text{adv}}$, the modality-common subspaces are not well blended, which indicates that the model cannot capture the shared semantics between the modalities.
- After removing $\mathscr{L}_{\text{rec}}$, the distributions of the modality-specific and modality-common subspace become unclear. Specifically, there exists discrepancies between common representations from different modalities; and commonality between specific representations. This proves that simple adversarial consistency constraints and difference constraints cannot guarantee the purity of modal disentanglement representations. This proves that simple difference and adversarial similarity constraints cannot effectively separate modality-common semantics from modality-specific semantics. And the reconstruction constraints can enhance the purity of the disentanglement representations.

5.4.3. *Effectiveness of each cross-modal clue*

To verify the validity of different cross-modal clues, we design the following variants:

- w/o $\mathbf{h}_{me}$: we remove the multimodal mutually enhanced representation $\mathbf{h}_{me}$ and only consider complementary and inconsistency clues.
- w/o $\mathbf{h}_{mi}$: we remove the multimodal inconsistency comparison representation $\mathbf{h}_{mi}$ and only consider enhancement and complementation clues.
- w/o $\mathbf{h}_t^s + \mathbf{h}_v^s$: we remove modality-specific representations $\mathbf{h}_t^s$ and $\mathbf{h}_v^s$ and only consider enhancement and inconsistency clues.

We compare the fake news detection performance of DCCMA-Net and its variants on three datasets. The comparison results are shown in Table 3. After removing any of the clues, the model performance declined. This suggests that the three cross-modal clues are effective in detecting fake news. Detecting different kinds of fake news needs to utilize distinct cross-modal clues. After removing any kind of cross-modal clues, the model loses the ability to detect the corresponding type of fake news, leading to decreased performance.

5.5. *Parameter sensitivity*

This section verifies the sensitivity of DCCMA-Net to hyperparameter settings.

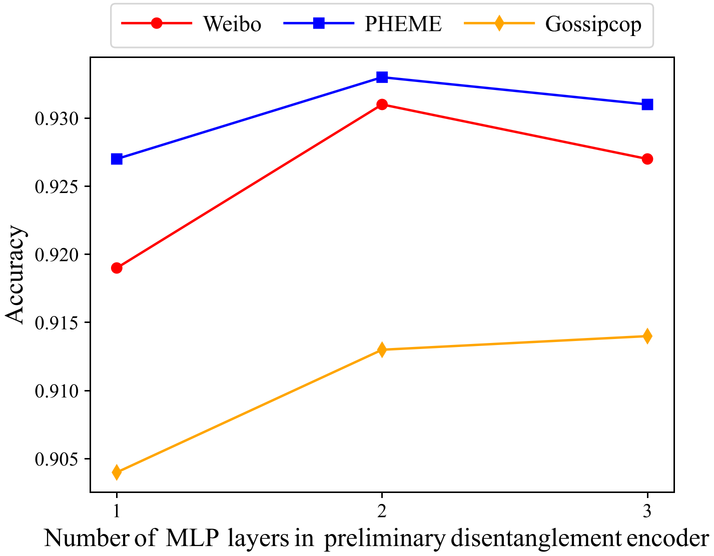


Fig. 12. The effect of different number of MLP layer in preliminary disentanglement encoder on model performance.

5.5.1. Effect of MLP layer numbers in preliminary disentanglement encoder to DCCMA-Net

Following the classical study of multimodal untwisted representation learning (Yang et al., 2022), DCCMA-Net uses a two-layer MLP to preliminarily decompose each modality into modality-common and modality-specific representations. In this subsection, we test the effect of the number of MLP layers on the DCCMA-Net. Specifically, we test the ACC of DCCMA-Net on three datasets for MLP layer numbers ranging from 1 to 3. The experimental results are shown in Fig. 12.

From the experimental results, it can be seen that when the number of MLP layers is 2, DCCMA-Net achieves good performance on all three datasets. When the number of MLP layers is 3, the performance of DCCMA-Net decreases on both Weibo and PHEME datasets, and only on the Gossipcop dataset has a slight improvement. This is because more parameters tend to cause overfitting of the model. However, the larger amount of data in the Gossipcop dataset mitigates the risk of overfitting. Experiments show that a preliminary disentanglement encoder consisting of two layers of MLPs can achieve the best performance.

5.5.2. Effect of disentanglement coefficient to DCCMA-Net

The total loss of the DCCMA-Net model is denoted as $\mathscr{L}_{\text{ALL}} = \mathscr{L}_{\text{CE}} + \lambda\mathscr{L}_{\text{DEC}}$. The objective of detecting fake news is fundamentally a classification problem. Hence the coefficient of the cross-entropy loss is assigned to 1. The hyperparameter λ controls the extent of the disentanglement loss. We perform experiments to test the sensitivity of DCCMA-Net to the disentanglement coefficient λ. As shown in Fig. 13, we test the performance of DCCMA-Net in the range of λ from 0.0 to 0.5. We can obtain the following conclusion from the experimental results:

Overall, DCCMA-Net achieves strong performance regardless of the value of λ, which proves that DCCMA-Net is not parameter-sensitive. In detail, the accuracy of DCCMA-Net obtains a substantial improvement when λ ranges from 0.0 to 0.1. This indicates that the disentangled representations do help in the fake news detection task. Specifically, the refined modality-specific and modality-common representations not only mitigate modal redundancy but also help cross-modal clues mining. When λ ranges from 0.1 to 0.3, the accuracy of DCCMA-Net continues to increase. But when λ goes from 0.1 to 0.3, the accuracy of DCCMA-Net decreases. This is because a moderate λ helps the model to capture modality-specific and modality-common representations that are useful for detection, but larger values of λ may distract the model from the main classification task. Therefore, we set the λ value to 0.3.

5.6. Case study

To visually demonstrate the effectiveness and explainability of our proposed DCCMA-Net model on fake news detection tasks, we select several representative cases of fake news from three datasets and record the contribution weights and predicted values that the DCCMA-Net outputs as shown in Fig. 14. Case 1 - Case 4 are from the Weibo dataset, Case 5 is from the PHEME dataset, and Case 6 is from the Gossipcop dataset.

We can observe that DCCMA-Net can reliably determine the veracity of the news and provide the corresponding weights of different clues. In case 1, the image and text semantics are consistent, but both overstate the event. Therefore, DCCMA-Net assigns the largest contribution weight to the modally-common representation. In case 2, the text is normal but the image is false. DCCMA-Net assigns the largest contribution weight to the text modality-specific representation. In case 3, the image describes a non-toxic mantis shrimp, but the text describes it as a highly toxic creature. The news text is false. The DCCMA-Net assigns the largest contribution weight to the text modality-specific and modality inconsistency representations. In case 4, the image depicts a famous Korean star,

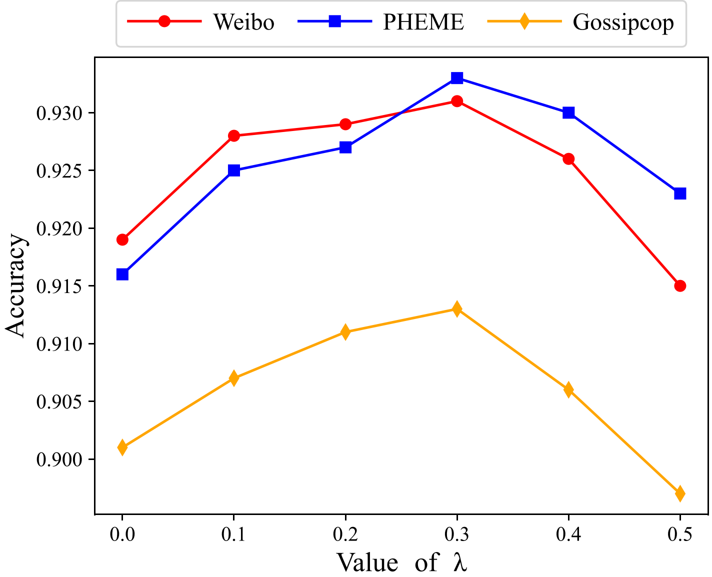


Fig. 13. The effect of different disentanglement coefficient λ on model performance.

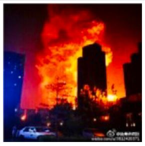
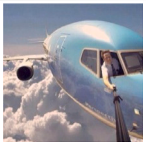
	Case 1:	Case 2:	Case 3:	Case 4:	Case 5:	Case 6:
Image						
Text	I believe thousands are dead with such huge explosion power. Some say it is fake, it is a rumor. They are fooled.	The pilot actually took a selfie in the air.	The water mantis lives in sewers. Its head has two to three times the poison of pufferfish and has no antidote.	Today, the Guangzhou Shahe on the morning of 3.15 has a terrorist incident.	NHL postpones Maple Leafs-Senators game after tragic shootings in Ottawa.	Angelina Jolie & Jared Leto Dating After Brad Pitt Divorce.
Contribution weight	$w_T = 0.21$ $w_V = 0.27$ $w_{ME} = 0.51$ $w_{MI} = 0.01$	$w_T = 0.01$ $w_V = 0.81$ $w_{ME} = 0.02$ $w_{MI} = 0.16$	$w_T = 0.55$ $w_V = 0.02$ $w_{ME} = 0.02$ $w_{MI} = 0.41$	$w_T = 0.34$ $w_V = 0.03$ $w_{ME} = 0.01$ $w_{MI} = 0.62$	$w_T = 0.18$ $w_V = 0.06$ $w_{ME} = 0.03$ $w_{MI} = 0.73$	$w_T = 0.16$ $w_V = 0.04$ $w_{ME} = 0.03$ $w_{MI} = 0.77$
Predict result	False	False	False	False	False	False
Ground truth	False	False	False	False	False	False

Fig. 14. The visualization of Contribution weights and predicted results captured by DCCMA-Net model.

while the text describes a terrorist incident. The image and the text are grossly inconsistent. DCCMA-Net assigns the greatest weight to the modality inconsistency representation. In case 5, the image and text are semantically inconsistent. Therefore DCCMA-Net assigns the maximum contribution weight to the modality inconsistent representation. The news story shown in case 6 is rumors about a dating relationship between two stars. Since there are no actual pictures of the date, the fake news creator splices two single photos as the dating photo. DCCMA-Net identifies semantic inconsistencies between the text and the fake image and assigns the maximum contribution weight to the modality inconsistency representation. The above case studies prove that DCCMA-Net not only accurately detects fake news, but also provides explanations for the detection results.

6. Discussion

6.1. Theoretical implications

In this paper, we propose a novel model, DCCMA-Net, which uses disentanglement representations to mine cross-modal clues and aggregate them for multimodal fake news detection. Most multimodal fake news detection studies do not distinguish between modality-specific semantics and modality-common semantics but consider each modality as a coupled whole. As a result, previous studies are not able to mine cross-modal clues explicitly and are at the black-box level. Unlike previous studies, DCCMA-Net disentangles each modality into modality-specific and modality-common semantics and utilizes these disentangled representations to explicitly and comprehensively mine cross-modal clues. In addition, DCCMA-Net proposes an adaptive attention aggregation mechanism that aggregates different clues by contribution to obtain highly discriminative representations and highlights the most contributive clues to provide explanations for the detection results.

6.2. Practical implications

Our proposed DCCMA-Net focuses on improving the performance and explainability for multimodal fake news detection, which has strong practical implications. First, DCCMA-Net can be applied to online social media platforms to detect fake news automatically. Previously, social media platforms often manually detect fake news, which is time-consuming and difficult to detect fake news in real-time. While DCCMA-Net can detect fake news automatically and effectively. So it can mitigate the negative impact of fake news. Second, DCCMA-Net can be used to provide explanations for detecting results. An explainable fake news detection system can enhance public trust. Finally, providing explanations for the detection results can help engineers retrace the model inference process, which can help identify and improve the weaknesses of the model.

6.3. Future work

In this subsection, we describe our future research directions as follows:

- **Detect fake news videos.** With the development of multimedia social networks, short video platforms such as YouTube[8] and TikTok[9] are gradually becoming the mainstream medium for information dissemination. News on short video platforms usually contains images, text, video, and audio, which has stronger propagation ability. Therefore, detecting fake news videos is highly significant. DCCMA-Net is used to detect the authenticity of news containing images and text. However, we believe that DCCMA-Net has the potential to detect fake news videos. The reason is that the core innovation of DCCMA-Net is to explicitly mine rich semantic associations between different modalities, which is modality-independent. However, DCCMA-Net cannot capture video and audio features, which is the major challenge for DCCMA-Net to detect fake news videos. In future work, we will extend DCCMA-Net to enable it to detect fake news videos.
- **Combine with large (vision) language models.** The Large Language Models (LLMs) are artificial intelligence models pre-trained on large-scale corpus (Li et al., 2024). LLMs have strong natural language understanding and generation capabilities. Large Visual Language Models (LVLMs) add visual information processing capabilities to LLMs and can process multimodal data well (Caffagni et al., 2024). Benefiting from the powerful capabilities of LLMs and LVLMs, we will explore the combination of DCCMA-Net with L(V)LMs in our future work. Specifically, we will combine L(V)LMs in the following three directions. First, **L(V)LMs act as background knowledge providers.** News on social networks tends to be shorter in length and have fewer images. Therefore, news often lacks the introduction of background knowledge (e.g., people, places, and events, etc.). Benefiting from the rich knowledge contained in L(V)LMs, we can try to let L(V)LMs provide the introduction of background knowledge for news. Specifically, let LLMs and LVLMs generate text descriptions and image captions to supplement the background knowledge of the original news. This can help DCCMA-Net capture more abundant news semantics. Secondly, **L(V)LMs act as tool users.** DCCMA-Net can mine multimodal features effectively for fake news detection. However, DCCMA-Net performs weakly in retrieval and verification of factual knowledge. LVLMs show good ability in integrating factual knowledge using retrieval-based tools, which can help DCCMA-Net to improve its factual verification capability. Third, **L(V)LMs act as explanation generators.** DCCMA-Net outputs the contribution weights of different cross-modal clues as explanations. Although this can provide explainability for fake news detection, this interpretation is not intuitive enough. L(V)LMs can generate easy-to-read and fluent natural language, which could help fill that gap. In future research, we will combine the cross-modal clues mining capability of DCCMA-Net and the natural language generation capability of L(V)LMs to generate easy-to-read natural language as explanations for multimodal fake news detection.

7. Conclusion

In this paper, we summarize three cross-modal clues that are effective for multimodal fake news detection: modality-common semantic enhancement, modality-specific semantic inconsistency, and modality-specific semantic complimentary. And we propose a unified framework, DCCMA-Net, to mine and aggregate these cross-modal clues to enhance the performance and explainability of multimodal fake news detection. DCCMA-Net disentangles each modality into refined modality-specific and modality-common representations and uses these disentangled representations to explicitly and comprehensively mine cross-modal clues. Afterward, DCCMA-Net reweights and aggregates these clues to obtain highly discriminative news representations and highlights the most contributive evidence. Extensive experiments on three public datasets demonstrate the performance and explainability of DCCMA-Net.

[8] https://www.youtube.com.

[9] https://www.tiktok.com.

CRediT authorship contribution statement

Siqi Wei: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Data curation, Conceptualization. **Zheng Wang:** Writing – review & editing, Validation, Software, Resources, Data curation. **Meiling Li:** Writing – review & editing, Validation, Software. **Xuanning Liu:** Writing – review & editing, Software. **Bin Wu:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the NSFC-General Technology Basic Research Joint Funds (grant no. U1936220) and the National Natural Science Foundation of China (grant no. 61972047, 62372060).

Data availability

Data will be made available on request.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., et al. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Alam, F., Cresci, S., et al. (2022). A survey on multimodal disinformation detection. In *Proceedings of the COLING* (pp. 6625–6643).

Bondielli, A., & Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. *Information Sciences*, *497*, 38–55.

Caffagni, D., Cocchi, F., Barsellotti, L., Moratelli, N., Sarto, S., Baraldi, L., et al. (2024). The revolution of multimodal large language models: A survey. In *Proceedings of the ACL* (pp. 13590–13618).

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. In *Proceedings of the WWW* (pp. 675–684).

Chen, S., Gong, C., Yang, J., Li, X., Wei, Y., & Li, J. (2018). Adversarial metric learning. In *Proceedings of the IJCAI* (pp. 2021–2027).

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the ICML* (pp. 1597–1607).

Chen, Y., Li, D., Zhang, P., Sui, J., Lv, Q., Tun, L., et al. (2022). Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the WWW* (pp. 2897–2905).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the ICLR* (pp. 1–21).

Duan, Y., Zheng, W., Lin, X., Lu, J., & Zhou, J. (2018). Deep adversarial metric learning. In *Proceedings of the CVPR* (pp. 2780–2789).

Guo, H., Zeng, W., Tang, J., & Zhao, X. (2023). Interpretable fake news detection with graph evidence. In *Proceedings of the CIKM* (pp. 659–668).

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415.

Hu, B., Sheng, Q., Cao, J., Shi, Y., Li, Y., Wang, D., et al. (2024). Bad actor, good advisor: Exploring the role of large language models in fake news detection. *vol. 38*, In *Proceedings of the AAAI* (pp. 22105–22113). 20.

Hu, L., Wei, S., Zhao, Z., & Wu, B. (2022). Deep learning for fake news detection: A comprehensive survey. *AI Open*, *3*, 133–155.

Jiang, G., Liu, S., Zhao, Y., Sun, Y., & Zhang, M. (2022). Fake news detection via knowledgeable prompt learning. *Information Processing & Management*, *59*(5), Article 103029.

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 ACM MM* (pp. 795–816).

Jing, J., Wu, H., Sun, J., Fang, X., & Zhang, H. (2023). Multimodal fake news detection via progressive fusion networks. *Information Processing & Management*, *60*(1), Article 103120.

Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications*, *80*(8), 11765–11788.

Kenton, J. D. M.-W. C., & Toutanova, L. K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the NAACL-HLT* (pp. 4171–4186).

Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019). MVAE: Multimodal variational autoencoder for fake news detection. In *Proceedings of the WWW* (pp. 2915–2921).

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the ICLR* (pp. 1–15).

Kwon, S., Cha, M., Jung, K., Chen, W., & Wang, Y. (2013). Prominent features of rumor propagation in online social media. In *Proceedings of the ICDM* (pp. 1103–1108).

Li, J., Yang, Y., Bai, Y., Zhou, X., Li, Y., Sun, H., et al. (2024). Fundamental capabilities of large language models and their applications in domain scenarios: A survey. In *Proceedings of the ACL* (pp. 11116–11141).

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), 1–35.

Luvembe, A. M., Li, W., Li, S., Liu, F., & Xu, G. (2023). Dual emotion based fake news detection: A deep attention-weight update approach. *Information Processing & Management*, *60*(4), Article 103354.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K. F., et al. (2016). Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the IJCAI* (pp. 3818–3824).

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11).

OpenAI (2023). ChatGPT: Optimizing language models for dialogue.. https://Openai.Com/Blog/Chatgpt/.

Popat, K., Mukherjee, S., Strötgen, J., & Weikum, G. (2016). Credibility assessment of textual claims on the web. In *Proceedings of the CIKM* (pp. 2173–2178).

Qi, P., Cao, J., Li, X., Liu, H., Sheng, Q., Mi, X., et al. (2021). Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues. In *Proceedings of the ACM MM* (pp. 1212–1220).

Qian, S., Wang, J., Hu, J., Fang, Q., & Xu, C. (2021). Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the SIGIR* (pp. 153–162).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the ICML* (pp. 8748–8763).

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(1), 5485–5551.

Reis, J. C., Correia, A., Murai, F., Veloso, A., & Benevenuto, F. (2019). Explainable machine learning for fake news detection. In *Proceedings of the WWW* (pp. 17–26).

Samadi, M., Mousavian, M., & Momtazi, S. (2021). Deep contextualized text representation and learning for fake news detection. *Information Processing & Management*, *58*(6), Article 102723.

Shang, L., Kou, Z., Zhang, Y., & Wang, D. (2022). A duo-generative approach to explainable multimodal COVID-19 misinformation detection. In *Proceedings of the WWW* (pp. 3623–3631).

Shu, K., Cui, L., Wang, S., Lee, D., & Liu, H. (2019). Defend: Explainable fake news detection. In *Proceedings of the ACM SIGKDD* (pp. 395–405).

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, *8*(3), 171–188.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *Proceedings of the ICLR*.

Singhal, S., Kabra, A., Sharma, M., Shah, R. R., Chakraborty, T., & Kumaraguru, P. (2020). Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). In *Proceedings of the AAAI* (pp. 13915–13916).

Singhal, S., Pandey, T., Mrig, S., Shah, R. R., & Kumaraguru, P. (2022). Leveraging intra and inter modality relationship for multimodal fake news detection. In *Proceedings of the WWW* (pp. 726–734).

Singhal, S., Shah, R. R., Chakraborty, T., Kumaraguru, P., & Satoh, S. (2019). Spotfake: A multi-modal framework for fake news detection. In *Proceedings of the bigdata* (pp. 39–47).

Song, C., Ning, N., Zhang, Y., & Wu, B. (2021). A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Information Processing & Management*, *58*(1), Article 102437.

Vaibhav, R. M. A., & Hovy, E. (2019). Do sentence interactions matter? Leveraging sentence level representations for fake news classification. In *Proceedings of the EMNLP* (pp. 134–139).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Proceedings of the neurIPS* (pp. 5998–6008).

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., et al. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, *18*(6), 1–26.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., et al. (2018). EANN: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the SIGKDD* (pp. 849–857).

Wang, B., Ma, J., Lin, H., Yang, Z., Yang, R., Tian, Y., et al. (2024). Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the WWW* (pp. 2452–2463).

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 ACM MM* (pp. 5696–5704).

Wani, A., Joshi, I., Khandve, S., Wagh, V., & Joshi, R. (2021). Evaluating deep learning approaches for covid19 fake news detection. In *Proceedings of the workshop on combating online hostile posts in regional languages during emergency situations* (pp. 153–163).

Wei, Z., Pan, H., Qiao, L., Niu, X., Dong, P., & Li, D. (2022). Cross-modal knowledge distillation in multi-modal fake news detection. In *Proceedings of the ICASSP* (pp. 4733–4737).

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the EMNLP* (pp. 38–45).

Wu, W., Wu, H., Jiang, L., Liu, X., Hong, J., Zhao, H., et al. (2024). From role-play to drama-interaction: An LLM solution. In *Proceedings of the ACL* (pp. 3271–3290).

Wu, Y., Zhan, P., Zhang, Y., Wang, L., & Xu, Z. (2021). Multimodal fusion with co-attention networks for fake news detection. In *Proceedings of the ACL* (pp. 2560–2569).

Xue, J., Wang, Y., Tian, Y., Li, Y., Shi, L., & Wei, L. (2021). Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management*, *58*(5), Article 102610.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In *Proceedings of the (neurIPS* (pp. 1–18).

Yang, D., Huang, S., Kuang, H., Du, Y., & Zhang, L. (2022). Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM international conference on multimedia* (pp. 1642–1651).

Yang, F., Liu, Y., Yu, X., & Yang, M. (2012). Automatic detection of rumor on sina weibo. In *Proceedings of the SIGKDD* (pp. 1–7).

Ying, Q., Hu, X., Zhou, Y., Qian, Z., Zeng, D., & Ge, S. (2023). Bootstrapping multi-view representations for fake news detection. In *Proceedings of the AAAI* (pp. 5384–5392).

Yu, F., Liu, Q., Wu, S., Wang, L., Tan, T., et al. (2017). A convolutional approach for misinformation identification.. In *Proceedings of the IJCAI* (pp. 3901–3907).

Zhang, X., Cao, J., Li, X., Sheng, Q., Zhong, L., & Shu, K. (2021). Mining dual emotion for fake news detection. In *Proceedings of the WWW* (pp. 3465–3476).

Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, *57*(2), Article 102025.

Zhang, W., Gui, L., & He, Y. (2021). Supervised contrastive learning for multimodal unreliable news detection in COVID-19 pandemic. In *Proceedings of the CIKM* (pp. 3637–3641).

Zhou, X., Wu, J., & Zafarani, R. (2020). Safe: Similarity-aware multi-modal fake news detection. In *Proceedings of the PAKDD* (pp. 354–367).

Zhou, Y., Yang, Y., Ying, Q., Qian, Z., & Zhang, X. (2023a). Multi-modal fake news detection on social media via multi-grained information fusion. In *Proceedings of the ICMR* (pp. 2295–2304).

Zhou, Y., Yang, Y., Ying, Q., Qian, Z., & Zhang, X. (2023b). Multimodal fake news detection via clip-guided learning. In *Proceedings of the ICME* (pp. 2825–2830).

Zhou, X., Zafarani, R., Shu, K., & Liu, H. (2019). Fake news: Fundamental theories, detection strategies and challenges. In *Proceedings of the WSDM* (pp. 836–837).

Zubiaga, A., Liakata, M., & Procter, R. (2017). Exploiting context for rumour detection in social media. *Lecture Notes in Computer Science*, 109–123.