Pattern Recognition 164 (2025) 111485

Contents lists available at ScienceDirect

Pattern Recognition

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

Balanced Multi-modal Learning with Hierarchical Fusion for Fake News Detection

Fei Wu [a,*], Shu Chen [a], Guangwei Gao [a], Yimu Ji [a,b], Xiao-Yuan Jing [c,d]

[a] College of Automation & College of Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing, 210023, China
[b] The State Key Laboratory of Blockchain and Data Security, Zhejiang University, Hangzhou, 310027, China
[c] Guangdong Provincial Key Laboratory of Petrochemical Equipment Intelligent Security, Guangdong University of Petrochemical Technology, Maoming, 525000, China
[d] School of Computer Science, Wuhan University, Wuhan, 430072, China

ARTICLE INFO

Keywords:
Fake news detection
Imbalanced multi-modal learning
Hierarchical Fusion

ABSTRACT

Multi-modal fake news detection (MFND) leverages data from various modalities, including text, image, video, and audio, to identify the authenticity of news content. Most existing MFND methods focus on extracting feature representations of each modality and integrating them by fusion strategies. However, they ignore the problem of modality imbalance where the dominant modality suppresses the performance of other modalities during optimization process, which leads to insufficient utilization of multi-modal information. To address the issue of modality imbalance and guarantee the effective utilization of each modality, we propose an approach called Balanced Multi-modal Learning with Hierarchical Fusion (BMLHF), which contains a Multi-modal Information Balancing (MIB) module and a Hierarchical Fusion (HF) module. Specifically, we extract multi-view semantic and pattern features of text and image. MIB calculates the modal information firstly to estimate the modal difference ratio, and it dynamically allocates corresponding weight for optimization of each view of modalities, which facilitates the modal information balance state. HF fully explores the diversity and correlation of multi-modal information in two stages. Intra-modal multi-view information fusion stage designs multi-view attention sub-network to sufficiently fuse semantic and pattern features within modalities. Inter-modal correlation fusion stage designs the joint correlation matrix based cross-attention strategy to learn multi-modal fused features with complementary characteristics. Extensive benchmark experiments demonstrate that our approach significantly surpasses state-of-the-art MFND methods.

1. Introduction

With the rapid development of information technology, online social media platforms such as Twitter [1] and Weibo [2] are becoming increasingly popular [3]. However, this rapid growth also brings about the inevitable spread of false information and fake news [4], including multiple modalities such as text, image, video, and audio. The dissemination of fake news has detrimental effects and causes significant negative influence [5]. Therefore, it is imperative to effectively discern the authenticity of multi-modal information shared on social media, namely multi-modal fake news detection (MFND) [6].

Currently, the research in the field of fake news detection has made great progress, which is mainly divided into two types: uni-modal methods and multi-modal methods. Uni-modal fake news detection methods can be divided into textual methods and visual methods. Specifically, current textual methods attempt to capture semantic [7], emotional [8], position-based [9] and intent-based [10] features from the perspective of textual content, as well as features such as comments [11], communication structures [12], and user profiles [13] from the perspective of social context, which have achieved relatively decent performance. For example, Hu et al. [14] designed a novel adaptive rationale guidance network that complements small and large LMs by selectively acquiring insights from LLM-generated rationales for small language models. Visual methods often use visual information such as images and videos. They usually use visual feature extractor to obtain visual features. For example, Sharif et al. [15] extracted image features with the pre-trained CNN. Simonyan et al. [16] obtained generic visual representations using VGG19.

Multi-modal fake news detection methods mainly face the modality differences among different modalities. Most existing MFND methods attempt to reduce the discrepancy between modalities to obtain discriminative features by interacting and fusing information between

[*] Corresponding author.
E-mail address: feiw@njupt.edu.cn (F. Wu).

https://doi.org/10.1016/j.patcog.2025.111485
Received 31 August 2024; Received in revised form 27 December 2024; Accepted 18 February 2025
Available online 28 February 2025

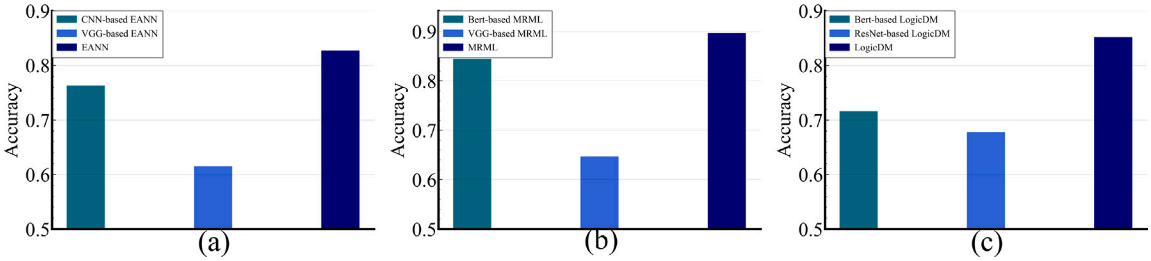


Fig. 1. Overall accuracy between uni-modal versions and multi-modal version on Weibo. (a) Text-only EANN, image-only EANN, and EANN. (b) Text-only MRML, image-only MRML, and MRML. (c) Text-only LogicDM, image-only LogicDM, and LogicDM.

multiple modalities. For example, Wang et al. [17] used event discriminator to extract event-invariant features. Singhal et al. [18] connected the pre-trained language features with the uni-modal features extracted from the visual modality. Zhou et al. [19] designed a similarity perception method to learn multi-modal features. Chen et al. [20] designed a cross-modal ambiguity learning module to estimate the ambiguity between different modalities. Peng et al. [21] designed triplet learning and contrastive pairwise learning to discover and capture the relationships within and between modalities. Liu et al. [22] proposed an interpretable multi-modal error information detection model based on neural symbolic. Ying et al. [23] proposed single-view prediction and cross-modal consistency learning to distinguish information in uni-modal and multi-modal features.

1.1. Motivations

Unfortunately, existing methods ignore the fact that the news information of each modality has a different influence on the detection task, i.e., the modality imbalance problem [24]. Currently, the methods solving modality imbalance problem focus on the field of audio-video classification. Among them, the prevailing practices include uni-modal assistance [25], gradient blending [26], and sample-level modal evaluation [27]. For example, Du et al. [25] strengthened the multi-modal model with the help of well-trained uni-modal models. Wang et al. [26] adopted gradient blending to obtain an optimal blending of modalities. Wei et al. [25] enhanced low-contribution modalities using sample-level modal evaluation metric. However, these methods only focus on solving the audio-video classification problem and do not consider the modality imbalance in MFND. News data has significant differences from audio-video data in terms of feature differences, modal correlation and fusion strategies, which makes these imbalanced multi-modal learning methods cannot be directly used in MFND.

To observe the impact of different modalities on the multi-modal model's accuracy, we selected three representative multi-modal methods (EANN [17], MRML [21], LogicDM [22]) and examined the accuracy results of their multi-modal and corresponding uni-modal versions. For text-only versions, EANN, MRML, and LogicDM separately use Text-CNN, BERT, and BERT. For image-only versions, they separately use VGG, VGG, and Resnet. Text-only version, image-only version, and complete version of EANN, MRML, and LogicDM adopt original setting in [17,21,22] for training with optimal hyperparameters on the Weibo dataset [2]. For testing samples on Weibo, we obtain corresponding features based on the trained models with optimal parameters, and then these features are fed into the classifier used in the original papers [17,21,22] to obtain the classification results. Fig. 1 shows the overall accuracy of models between uni-modal versions and the complete multi-modal version on the Weibo dataset [2]. From the figure, we can find that the accuracy is much higher for text modality than image modality for these three models, which means text modality plays a much more important role than image modality for MFND.

Fig. 2 shows the batch-average uni-modal logit score of EANN, MRML, and LogicDM, where the logit score is the output of the last fully connected layer. We use the logit score as modal information in this paper. "Text-multi" and "Image-multi" represent uni-modal logit scores in multi-modal model, and the "Image-only" represents the logit score of single-image-modality model. Fig. 2 indicates that in the progress of training iteration, the logit score of the text modality increases while the image one tends to flatten out. From Fig. 2, we can find that the logits of the "Image-only" are significantly higher than those of "Image-multi" during the training process. During multi-modal joint training, the logits of "Image-multi" are significantly lower than those of "Text-multi", which proves that the modality with low contribution is suppressed during the training process.

Moreover, [28] considered that, if the logit of a certain modality is small, it may indicate that model encounters difficulties in extracting and utilizing the information from this modality. From finding of [28] and Figs. 1, 2, we believe that it is the imbalance reflected in the logit score that leads to the inadequate model optimization that results in the decline of model accuracy. In other words, the dominant modality restrains the performance of the other modalities, and the information of other modalities is not fully utilized, leading to the phenomenon of modality imbalance.

For multi-modal fake news detection task, there is a lack of methods specifically aimed at addressing the issue of modality imbalance.

1.2. Contributions

To solve the modality imbalance problem and improve the discriminability of multi-modal fused features, we propose an approach called Balanced Multi-modal Learning with Hierarchical Fusion (BMLHF), which includes a Multi-modal Information Balancing (MIB) module and a Hierarchical Fusion (HF) module. Specifically, MIB utilizes a four-channel network tailored for both images and text, extracting semantic and pattern features as multi-view features from two modalities. MIB dynamically assigns the weight for optimization of each view of modalities to balance modal information. HF designs the multi-view attention network to fuse semantic and pattern features within modalities, and then uses the designed joint correlation matrix based cross-attention strategy to mine complementarity information between modalities.

The main contributions this paper are as follows.

(1) We design a Multi-modal Information Balancing (MIB) module, which calculates the modal information firstly, and it obtains the modal difference ratio and dynamically allocates corresponding weights to different views of modalities during their optimization process, thereby adaptively adjusting the optimization process of each modality to achieve modal information balance. To the best of our knowledge, our approach is a relatively early work in investigating the modality imbalance problem in MFND.

(2) We additionally propose a Hierarchical Fusion (HF) module to perform cross-modal and cross-view fusion, including a dual cross-transformer interaction block and two fusion stages, which fully considers the correlation and importance of different modalities. The dual cross-transformer interaction block facilitates the interaction of information from different modalities. Intra-modal multi-view information fusion stage fuses multi-view features within modalities. Inter-modal correlation fusion stage excavates complementary information between

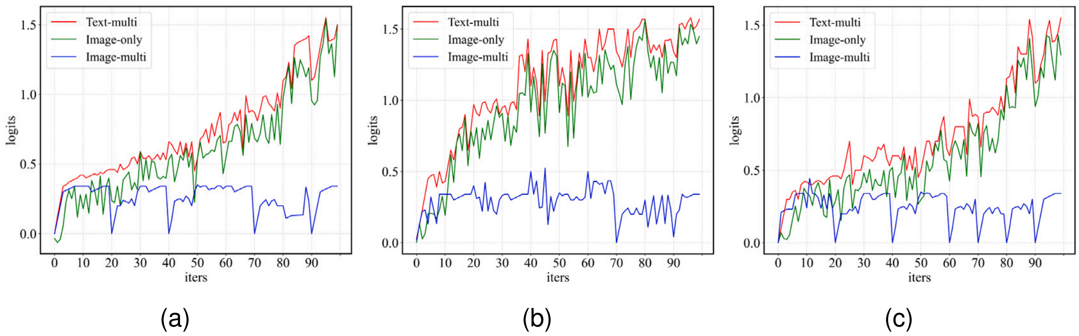


Fig. 2. The batch-average uni-modal logit scores on Weibo. (a) EANN, (b) MRML, (c) LogicDM.

modalities.

(3) Experimental results on Twitter [1], Weibo [2] and Fakeddit [29] demonstrate that the proposed approach outperforms the state-of-the-art MFND methods by a large margin.

1.3. Organization

The remaining content of this paper is structured as follows. Introductions of fake news detection methods and imbalanced multi-modal learning methods are reviewed in Section 2. Section 3 illustrates the framework of BMLHF and design of each module in detail. Experiments and settings are studied in Section 4. Finally, Section 5 concludes our work and the effectiveness of our model.

2. Related works

The task of fake news detection is a prevalent and critical topic in real life [30–32]. Due to the development of news data, fake news detection can be divided into uni-modal fake news detection and multi-modal fake news detection.

2.1. Uni-modal fake news detection

The purpose of uni-modal fake news detection is to improve the ability of fake news detection by using textual information or visual information.

(1) **Textual news.** Textual news is usually processed as text embeddings derived at the word, sentence, and document levels. In this context, a news article can be represented by latent vectors, which can either be utilized as input for classifiers directly or used in various network models [30]. Ma et al. [33] used RNN as the basic model and captured the relevant information of the event over time by learning its hidden-layer representations. With the development of graph learning, Vaibhav et al. [34] modeled each news article as a graph and redefined the fake news detection task as a graph classification task, where the nodes represent the sentences of the article and the edges represent the semantic similarity between pairs of sentences. Giachanou et al. [35] considered the role of emotional signals and proposed a LSTM model that incorporates emotional signals obtained from the text of the claims in order to distinguish between real and fake news.

(2) **Visual news.** Some early studies utilized the basic statistical features of images [36,37], such as the number of images, image visibility and image type [38], to help detect fake news. Researchers [39] extracted advanced image features and combined them with post-based and user-based features to discover fake news. However, these features cannot adequately represent image features at the complex visual level. Inspired by the capability of CNN, several existing efforts [17, 40] obtained generic visual representations using the pre-trained deep CNN such as VGG19 [16]. In order to better utilize the intrinsic features of fake news images and other task-relevant information, Qi et al. [41] proposed a multi-domain visual neural framework that combines frequency-domain and pixel-domain visual information to distinguish real news from fake ones by visual features. This method automatically captures image features in the frequency domain using a CNN and automatically extracts image semantic features in the pixel domain using a CNN-RNN architecture.

2.2. Multi-modal fake news detection

The usage of pure textual information and visual information is effective for fake news detection, and it is feasible to consider them together [42–44]. Earlier studies [16,17] believed that visual information is a compensation for text information, and thus they used visual extractors to extract visual features and splice them with text features. Generally, multi-modal models obtain image features from pre-trained VGG19 [16] first for fake news detection, and then concatenate these visual features simply with textual features. However, it has not fully considered the difference between visual and textual information. Therefore, some researchers propose that textual and visual news are related at a high-level semantic level and should not fuse features in a coarse-grained manner. Qian et al. [45] proposed the Hierarchical Multi-modal Contextual Attention Network (HMCAN) architecture to jointly consider the multi-modal context information and hierarchical semantics of text in a deep, unified framework. Singhal et al. [18] connected the pre-trained language features with the single modality features extracted from the visual modality. Zhou et al. [19] designed a similarity perception method to learn multi-modal features. Chen et al. [20] designed a cross-modal ambiguity learning module to estimate the ambiguity between different modalities. Peng et al. [21] designed triplet learning and contrastive pair learning to discover and capture the relationships within and between modalities. Liu et al. [22] proposed an interpretable multi-modal misinformation detection model based on neural-symbolic. Ying et al. [23] proposed a single-view prediction and cross-modal consistency learning method to distinguish information in uni-modal and multi-modal features. Yu et al. [46] integrated multi-modal features and passed them through a quantum convolutional neural network (QCNN) to obtain discriminative results.

However, existing methods focus on the differences between the modalities, but they overlook the modality imbalance issue.

2.3. Imbalanced multi-modal learning

In real life, multi-modal data is imbalanced, and early studies [24] have shown that the dominant modality will undoubtedly have an inhibitory effect on other modalities. It is the priority for imbalanced multi-modal learning methods to weaken this inhibitory effect and balance the information between modalities. According to recent studies [26,28], multi-modal models that optimize a uniform learning objective for all modalities with a joint training strategy will be inferior to uni-modal models in some situations. Such a phenomenon

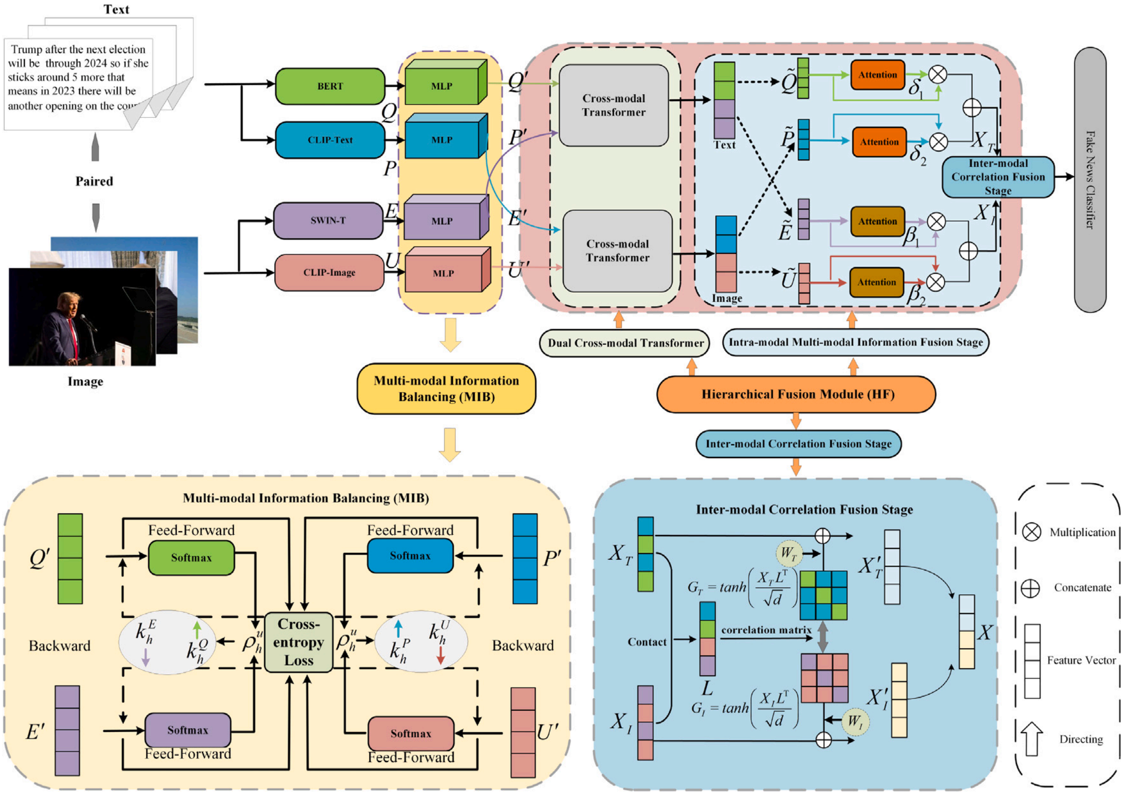


Fig. 3. The architecture of our BMLHF model.

contradicts the intention of improving model performance through integrating information from multiple modalities.

Previous researchers on audio-visual classification [47] claimed that various modalities tend to converge at different rates, leading to an uncoordinated convergence problem. To cope with this problem, some methods aid the training of multi-modal models with the help of additional uni-modal classifiers or pre-trained model. Du et al. [25] trained the multi-modal model by distilling knowledge from well-trained uni-modal models to strengthen the multi-modal model. However, it inevitably requires extra effort to train additional neural components. Some recent researches propose that modality imbalance is caused by the discrepancy in the model update process between modalities. Wang et al. [47] first proposed gradient blending to obtain an optimal blending of modalities based on their over-fitting behaviors. Wu et al. [26] tackled the problem by adaptively controlling the optimization rate of each modality. Xu et al. [28] proposed a plug-and-play multi-modal cosine loss to achieve a more balanced optimization process. Wei et al. [27] proposed a sample-level modal evaluation metric method for enhancing low-contribution modalities and reasonably improving multi-modal cooperation.

To sum up, existing imbalanced multi-modal learning methods are concentrated on audio-visual classification, and cannot be directly used for news detection task. For multi-modal fake news detection methods, the research on imbalanced multi-modal learning has not been well studied.

3. Proposed approach

We propose an approach called Balanced Multi-modal Learning with Hierarchical Fusion (BMLHF), which contains a Multi-modal Information Balancing (MIB) module to address the issue of modality imbalance and a Hierarchical Fusion (HF) module to capture discriminative fused news features. Fig. 3 shows the structure of BMLHF.

3.1. Joint semantic and pattern feature extraction module

Given the set of news $O = \{o_n\}_{n=1}^{N}$ includes text and image modalities, and the image modality is represented as $I = \{i_n\}_{n=1}^{N}$, the text modality is represented as $T = \{t_n\}_{n=1}^{N}$. The one-hot label matrix $Y = \{y_n\}_{n=1}^{N} \in \mathbb{R}^{M \times N}$ is corresponding to O, where M is the number of categories of the news. We aim to learn a mapping function: $O \rightarrow Y$. To fully exploit the information within each modality, we extract features from two perspectives: pattern and semantic features. We use BERT [48] and Swin-T [49] for text and image to extract pattern features which reveals the basic fine-grained characteristics of modalities. We use CLIP [50] for both text and image to extract semantic features which reveal the cross-modal semantic correspondence of modalities as a complement to pattern features to extract multi-view features.

BERT is the most popular pre-trained language model based on transformer in NLP. Swin-T is a vision transformer with hierarchical architecture which has made sensational progress in tasks like object detection. CLIP has already given due consideration to cross-modal correlation during the pre-training process [50], which is appropriate to take advantage of CLIP to dig deep cross-modal semantic information. Thus, BMLHF utilizes BERT, Swin-T, CLIP-text and CLIP-image as four channels for feature extraction.

Specifically, for text modality, we extract the pattern feature $Q = \{q_n\}_{n=1}^{N} \in \mathbb{R}^{d_t \times N}$ via BERT and semantic feature $P = \{p_n\}_{n=1}^{N} \in \mathbb{R}^{d_p \times N}$ with CLIP-text, where d_t, d_p represent dimensionality of q_n and p_n. For image modality, the pattern feature $E = \{e_n\}_{n=1}^{N} \in \mathbb{R}^{d_e \times N}$ and the semantic feature $U = \{u_n\}_{n=1}^{N} \in \mathbb{R}^{d_p \times N}$ are extracted by Swin-T and CLIP-image, where d_e, d_p represent the dimensionality of e_n and u_n. For each channel, these four types of features are projected to the same dimension with the two-layer MLP g^u, and we can obtain the encoded features $Q' = \{q'_n\}_{n=1}^{N} \in \mathbb{R}^{d_k \times N}$, $E' = \{e'_n\}_{n=1}^{N} \in \mathbb{R}^{d_k \times N}$, $P' = \{p'_n\}_{n=1}^{N} \in \mathbb{R}^{d_k \times N}$ and $U' = \{u'_n\}_{n=1}^{N} \in \mathbb{R}^{d_k \times N}$, where d_k represents the dimensionality of features. We define $Z = \left\{z_n^u\right\}_{n=1}^{N} \in \left\{Q', E', P', U'\right\}$, where $u \in \{Q, E, P, U\}$ represents a set of four types of features, i.e., four views.

3.2. Multi-modal information balancing (MIB) module

The problem of modality imbalance is a phenomenon where the dominant modality suppresses the performance of other modalities during optimization process. To deal with the problem, we design the MIB module to dynamically allocate weights to these modalities during their optimization process, adaptively regulating the model's optimization process to achieve modality balance.

For each channel of g^u, the parameter is θ^u, and the process of gradient updating is formulated by:

$$\theta^u_{h+1} = \theta^u_h - \lambda \tilde{g}^u(\theta^u_h) \tag{1}$$

where θ^u_h is the parameter after the hth iteration update.

We adjust the gradient of each modality adaptively by monitoring the amount of each modality's information. The logit score encapsulates the specific information of each modality, as evidenced by the fact that the logit score directly reflect the activation degree of different modalities. Therefore, in this paper, we calculate logit score of pattern and semantic views of each modality to investigate the modal information in more fine-grained details, which can be formulated as $W^u_h g^u(z^u_{i\,(h)}; \theta^u_h) + \frac{b_h}{4}$, where W^u_h and $\frac{b_h}{4}$ are the parameters of g^u, and $z^u_{i\,(h)}$ represents the ith feature of B_h, where B_h is a random batch in the hth iteration. We define s^u_i to further quantify the information of different views of modalities.

$$s^u_{i\,(h)} = \sum_{K=1}^{M} 1_{K=y_i} softmax(W^u_h g^u(z^u_{i\,(h)}; \theta^u_h) + \frac{b_h}{4})_K \tag{2}$$

where $y_i \in \{1, 2, \dots, M\}$, M is the number of categories.

We design the difference rate ρ^u_h to measure the influence of the modalities on the optimization process:

$$\rho^u_h = \frac{\sum_{i\in B_h} s^u_{i\,(h)}}{\sum_{i\in B_h}(s^Q_{i\ (h)} + s^E_{i\ (h)} + s^P_{i\ (h)} + s^U_{i\ (h)})} \tag{3}$$

The lager ρ^u_h is, the greater the amount of information in the corresponding view of modality compared to the other views of modalities, resulting in modality imbalance.

The modality with larger amount of information plays a dominant role in the optimization process of the model [26,28], and inhibits the optimization process of other modalities. Therefore, we design balance factors k^u_h to balance the optimization process of each modality:

$$k^u_h = \begin{cases} 1 - tanh(\alpha_1 \rho^u_h) & \rho^u_h > \frac{1}{4}, \\ 1 & others. \end{cases} \tag{4}$$

Then we integrate k^u_h into Eq. (1). The update process of θ^u_h is as follows:

$$\theta^u_{h+1} = \theta^u_h - \lambda k^u_h \tilde{g}^u(\theta^u_h) \tag{5}$$

By using k^u_h, we inhibit the optimization of the views with better performance ($\rho^u_h > \frac{1}{4}$), while the views with poor performance are not affected. Through our method, the optimization process of each modality can be modulated and the problem of modality imbalance can be effectively alleviated.

3.3. Hierarchical Fusion (HF) module

3.3.1. Dual cross-transformer interaction block

In this section, we design a dual cross-transformer interaction block to deal with the modality difference issue, facilitating subsequent multi-modal fusion. In particular, for pattern features Q' and E', we use $Cross{-}Transformer_{pattern}$ to implement the pattern interaction between them as follows:

$$\widetilde{Q}, \widetilde{E} = Cross - Transformer_{pattern}(Q', E'; \Theta_C) \tag{6}$$

Similarly, for semantic features P' and U', we have:

$$\widetilde{P}, \widetilde{U} = Cross - Transformer_{semantic}(P', U'; \Theta_D) \tag{7}$$

where Θ_C and Θ_D represent the parameters of the cross-modal transformers.

Our designed dual cross-modal transformer interacts information of different modalities from pattern and semantic aspects, generating richer and more comprehensive feature representations. This operation improves information consistency across modalities.

3.3.2. Two-stage fusion block

Effectively integrating pattern and semantic features can enrich and enhance the feature representation. Furthermore, multiple modalities convey different information, and it is necessary to effectively capture their complementary relationships. Therefore, for each modality, we design an intra-modal multi-view information fusion stage to fully exploit diverse information from the semantic and pattern aspects. For multi-modal fusion, we design an inter-modal correlation fusion stage to effectively obtain complementary information.

Intra-modal Multi-view Information Fusion Stage. In the first stage, we use the attention mechanism to obtain the fused features within the modality.

$$\delta_1 = f(\widetilde{Q}; \Theta_A), \delta_2 = f(\widetilde{P}; \Theta_A) \tag{8}$$

$$\beta_1 = f(\widetilde{E}; \Theta_B), \beta_2 = f(\widetilde{U}; \Theta_B) \tag{9}$$

where Θ_A, Θ_B are the parameters of the attention networks, δ_1, δ_2 are the attention coefficients of $\widetilde{Q}$ and $\widetilde{P}$, β_1, β_2 are the attention coefficients of $\widetilde{E}$ and $\widetilde{U}$. The fused features of the corresponding modality can be obtained by:

$$X_I = \delta_1 \widetilde{Q} + \delta_2 \widetilde{P} \tag{10}$$

$$X_T = \beta_1 \widetilde{E} + \beta_2 \widetilde{U} \tag{11}$$

where X_I and X_T separately represent the output features of image and text in the intra-modal multi-view information fusion stage.

Inter-modal Correlation Fusion Stage. In the second stage, we aim to obtain the fused features of different modalities. Due to the diverse information conveyed by multiple modalities, their correlation relationships need to be explored. Specifically, we design an inter-modal correlation fusion stage, and it first stitches the features of image and text modalities to obtain the joint representation $L = [X_I; X_T] \in R^{d\times N}$, where $d = 2d_k$. The joint correlation matrix G_I of image modality can be calculated as follows:

$$G_I = tanh(\frac{X_I L^{\mathrm{T}}}{\sqrt{d}}) \tag{12}$$

The correlation matrix G_T of text modality can be calculated similarly. G_I and G_T measure both inter-modal and intra-modal correlation relationships. The higher the correlation coefficient of paired features in G_I and G_T, the stronger the correlation of them within or between modalities. Therefore, the joint correlation matrices can characterize the complementary relationships of features in different modalities. We further learn the attention weight matrices W_I, W_T to calculate the joint correlation features X'_I, X'_T:

$$X'_I = ReLu(X_I + W_I G_I) \tag{13}$$

$$X'_T = ReLu(X_T + W_T G_T) \tag{14}$$

where X'_I and X'_T are further spliced to obtain final feature representation:

$$X = \{x_n\}_{n=1}^{N} = [X'_I; X'_T] \tag{15}$$

By using the hierarchical fusion module, semantic and pattern information is effectively integrated, enriching the feature representation of each modality. Moreover, the complementary information between different modalities is fully exploited, such that discriminative information of multiple modalities is fully utilized for detection task.

Algorithm 1 Balanced Multi-modal Learning with Hierarchical Fusion

```
Input data: Training set $O$, with text modality $T$ and image modality $I$, label matrix $Y$. Testing set $O'$, with text modality $T'$ and image modality $I'$.
for ($i = 1, \ldots, Y$) do
1. Calculate the modal information $s^u_{i(h)}$ by Eq. (2), $u \in \{Q, E, P, U\}$;
2. Calculate the difference rate $\rho^u_h$ by Eq. (3);
3. Calculate the balance factors $k^u_h$ by ;
4. Obtain final feature representation $X$ by Eq. (15);
5. Calculate $L_c$, $L_s$, $L_{total}$ by Eqs. (17), (18), and (19);
6. Updating $\theta^u_h$, $\Theta_A$, $\Theta_B$, $\Theta_C$, $\Theta_D$, $\theta_c$;
Output: Class label $\hat{Y}'$ of $O'$.
```

3.4. Fake news classifier

After obtaining the fused features of the image and text, we feed them into a classifier which contains a one-layer of MLP and $ReLu$ activation function to obtain the predicted label $\hat{y}_n$:

$$\hat{y}_n = C(x_n; \theta_c) \tag{16}$$

where $C(x_n; \theta_c)$ is a classifier, θ_c is the parameter of the classifier.

3.5. Total loss

In order to enhance the classification ability of the model, we minimize the cross-entropy loss L_c.

$$L_c = y_n log(\hat{y}_n) + (1 - y_n) log(1 - \hat{y}_n) \tag{17}$$

In order to enhance the discriminative ability of feature X, we employ a cross-modal supervised contrastive loss L_s. $P(n)$ is an index set of the same category as x_n in X, $A(n)$ is an index set of different categories from x_n in X, we can define L_s:

$$L_s = \sum_{n=1}^{N} \frac{-1}{|P(n)|} \sum_{p \in P(n)} log \frac{exp(x_n x_p/\tau)}{\sum_{a \in A(n)} exp(x_n x_p/\tau)} \tag{18}$$

$|P(n)|$ is the cardinality of $P(n)$, τ is a scalar parameter that is empirically set as $\tau = 0.5$. Thus, the total loss can be formulated as:

$$L_{total} = L_c + \alpha_2 L_s \tag{19}$$

where α_2 is a balance factor.

We use the Adam optimizer to pursue the optimal network parameters under the total loss. For the test sample set O', we use the optimal parameters Θ_A, Θ_B, Θ_C, Θ_D, θ_c, θ^u_h to obtain the corresponding features $\widetilde{O}$, and then we input these features into the classifier to obtain the label $\hat{Y}'$. The overall model optimization process of BMLHF is summarized as Algorithm 1.

4. Experiments

4.1. Datasets

We exploit three widely used datasets for fake news detection, i.e., Twitter [1], Weibo [2], and Fakeddit [29]. The details of these three datasets are as follows:

Weibo Dataset [2] is collected by Jin and is applied to the detection of multi-modal fake news. Fake news comes from the official rumor refutation platform, which is constructed through crowd-sourcing or official rumor refutation. We divide the dataset into a training set and a testing set with a ratio of 8:2, where the training set contains 7532 news items and the testing set contains 1996 news items.

Twitter Dataset [1] comes from the MediaEval Verifying Multimedia Use benchmark, which has also been applied to multi-modal fake news detection. We divide the dataset into a training set and a testing set, where the training set contains 8617 news items and the testing set contains 2059 news items.

Fakeddit Dataset [29] is a multi-modal benchmark dataset created by the research team at the University of California. It contains over one million samples of false and true information. These samples are classified into six categories, including true content, misleading content, etc. It encompasses multi-modal information such as text and image data. Following [51], we select 30,000 image–text pairs for training and 10,000 image–text pairs for testing.

4.2. Experimental setting

For textual representation, the embeddings of each token are obtained by pre-trained BERT-base model, and the embedding dimension is set to 768. For visual representation, we set the dimension size as 1000. In the joint semantic and pattern feature extraction module, the dimension of the hidden layer of MLP is 384, and the $ReLu$ activation function is used for MLP. The maximum training epoch number is set to 100. The learning rate is fixed to 0.001 for three datasets. We use Adam optimizer to train our model with a weight decay of 0.001. The number of heads of cross-transformer block is set to 4. We employ a grid search method to fine-tune the hyperparameters. We set τ, α_1, α_2 to 0.5, 0.8, 0.5 in the search range [0,1], which generate the best results.

4.3. Experimental results

4.3.1. Baselines

To evaluate the validity of BMLHF and to get a fair comparison, we select two types of competing baselines on datasets Twitter and Weibo, including uni-modal methods and multi-modal methods.

(1) Uni-modal baselines:

- **BERT** [48] is a popular pre-trained model for text. In experiment, we use BERT to extract textual features, followed by a classifier for fake news detection.
- **Swin-T** [49] is also an effective model to extract visual features. In experiment, we reserve only Swin-T to extract image features, followed by a classifier for fake news detection.

(2) Multi-modal baselines:

- **EANN** [17] uses event discriminator to extract event-invariant features.
- **SpotFake**+ [18] uses BERT and VGG to extract features from text and image, and then connects the corresponding features.
- **SAFE** [19] designs a similarity perception method to learn multi-modal features for fake news detection.
- **CAFE** [20] designs a cross-modal ambiguity learning module to estimate the ambiguity between different modalities, which endeavors to quantify the ambiguity between text and image for detection.
- **MRML** [21] designs triplet learning and contrastive pairwise learning to discover and capture the relationships within and between modalities.
- **LogicDM** [22] designs an interpretable multi-modal misinformation detection model based on neural symbolic AI.
- **BMR** [23] designs single-view prediction and cross-modal consistency learning to distinguish information in uni-modal and multi-modal features.
- **QMFND** [46] integrates multi-modal features and passes them through a quantum convolutional neural network (QCNN) to obtain discriminative results.
- **MTTV** [52] makes multi-modal data interact fully and capture the semantic relationships between them, and it proposes a scalable classifier to improve the classification balance of fine-grained fake news detection with the problem of class imbalance.

Table 1
Performance comparison between BMLHF and state-of-the-art methods on Weibo, Twitter, and Fakeddit datasets.

Dataset	Method	Acc	Fake news			Real news		
			Pre	Rec	F1	Pre	Rec	F1
Weibo	BERT	0.804 ± 0.022	0.800 ± 0.021	0.860 ± 0.012	0.830 ± 0.021	0.840 ± 0.012	0.760 ± 0.011	0.800 ± 0.030
	Swin-T	0.643 ± 0.014	0.640 ± 0.020	0.600 ± 0.018	0.620 ± 0.025	0.640 ± 0.014	0.700 ± 0.029	0.660 ± 0.010
	EANN	0.782 ± 0.020	0.827 ± 0.017	0.697 ± 0.034	0.756 ± 0.027	0.753 ± 0.018	0.863 ± 0.010	0.804 ± 0.006
	SpotFake+	0.870 ± 0.033	0.855 ± 0.032	0.892 ± 0.033	0.873 ± 0.032	0.769 ± 0.033	0.807 ± 0.032	0.787 ± 0.033
	SAFE	0.839 ± 0.015	0.840 ± 0.025	0.820 ± 0.015	0.830 ± 0.010	0.850 ± 0.015	0.830 ± 0.025	0.840 ± 0.010
	CAFE	0.840 ± 0.023	0.825 ± 0.036	0.851 ± 0.015	0.837 ± 0.026	0.855 ± 0.017	0.830 ± 0.024	0.842 ± 0.018
	MRML	0.897 ± 0.012	0.896 ± 0.012	0.905 ± 0.012	0.901 ± 0.010	0.898 ± 0.012	0.887 ± 0.010	0.892 ± 0.011
	LogicDM	0.852 ± 0.010	0.843 ± 0.015	0.859 ± 0.020	0.851 ± 0.015	0.862 ± 0.010	0.845 ± 0.012	0.853 ± 0.007
	BMR	0.831 ± 0.007	0.831 ± 0.024	0.824 ± 0.023	0.827 ± 0.022	0.831 ± 0.022	0.838 ± 0.023	0.834 ± 0.024
	QMFND	0.869 ± 0.016	0.900 ± 0.010	0.810 ± 0.015	0.850 ± 0.011	0.840 ± 0.011	0.920 ± 0.012	0.880 ± 0.011
	MTTV	0.876 ± 0.018	0.865 ± 0.018	0.897 ± 0.012	0.875 ± 0.012	0.890 ± 0.015	0.861 ± 0.018	0.875 ± 0.018
	BMLHF	**0.912** ± 0.009	0.930 ± 0.012	0.880 ± 0.010	**0.903** ± 0.013	0.894 ± 0.010	0.920 ± 0.009	**0.902** ± 0.012
Twitter	BERT	0.642 ± 0.012	0.602 ± 0.019	0.474 ± 0.009	0.526 ± 0.012	0.666 ± 0.008	0.766 ± 0.021	0.711 ± 0.012
	Swin-T	0.760 ± 0.015	0.720 ± 0.023	0.785 ± 0.029	0.740 ± 0.023	0.800 ± 0.021	0.753 ± 0.021	0.787 ± 0.022
	EANN	0.648 ± 0.016	0.810 ± 0.025	0.498 ± 0.040	0.617 ± 0.023	0.584 ± 0.016	0.759 ± 0.011	0.660 ± 0.027
	SpotFake+	0.790 ± 0.013	0.786 ± 0.014	0.747 ± 0.013	0.766 ± 0.014	0.793 ± 0.016	0.827 ± 0.011	0.810 ± 0.014
	SAFE	0.766 ± 0.036	0.752 ± 0.029	0.731 ± 0.020	0.742 ± 0.026	0.777 ± 0.029	0.795 ± 0.029	0.786 ± 0.026
	CAFE	0.806 ± 0.012	0.805 ± 0.011	0.813 ± 0.012	0.809 ± 0.011	0.807 ± 0.014	0.799 ± 0.014	0.803 ± 0.011
	MRML	0.803 ± 0.025	0.777 ± 0.025	0.747 ± 0.035	0.762 ± 0.030	0.821 ± 0.025	0.844 ± 0.030	0.832 ± 0.035
	LogicDM	0.911 ± 0.008	0.913 ± 0.016	0.958 ± 0.012	0.935 ± 0.010	0.909 ± 0.012	0.816 ± 0.025	0.859 ± 0.009
	BMR	0.872 ± 0.018	0.885 ± 0.014	0.931 ± 0.016	0.907 ± 0.016	0.842 ± 0.013	0.751 ± 0.012	0.794 ± 0.020
	QMFND	0.918 ± 0.015	0.880 ± 0.012	0.970 ± 0.008	0.920 ± 0.010	0.970 ± 0.015	0.870 ± 0.012	0.910 ± 0.008
	MTTV	0.885 ± 0.011	0.872 ± 0.023	0.912 ± 0.016	0.886 ± 0.008	0.910 ± 0.010	0.876 ± 0.016	0.886 ± 0.008
	BMLHF	**0.966** ± 0.013	0.985 ± 0.012	0.933 ± 0.010	**0.957** ± 0.010	0.948 ± 0.010	0.975 ± 0.012	**0.956** ± 0.010
Fakeddit	BERT	0.878 ± 0.010	0.853 ± 0.011	0.902 ± 0.013	0.878 ± 0.012	0.903 ± 0.010	0.861 ± 0.011	0.898 ± 0.012
	Swin-T	0.633 ± 0.023	0.503 ± 0.022	0.788 ± 0.012	0.663 ± 0.010	0.802 ± 0.023	0.651 ± 0.012	0.727 ± 0.012
	EANN	0.724 ± 0.020	0.727 ± 0.033	0.719 ± 0.014	0.723 ± 0.015	0.722 ± 0.014	0.729 ± 0.034	0.726 ± 0.017
	SpotFake+	0.819 ± 0.017	0.801 ± 0.018	0.848 ± 0.029	0.824 ± 0.020	0.839 ± 0.012	0.790 ± 0.021	0.813 ± 0.018
	SAFE	0.846 ± 0.011	0.809 ± 0.023	0.857 ± 0.010	0.832 ± 0.013	0.879 ± 0.014	0.837 ± 0.022	0.858 ± 0.012
	CAFE	0.912 ± 0.023	0.946 ± 0.019	0.886 ± 0.034	0.916 ± 0.023	0.878 ± 0.016	0.942 ± 0.018	0.909 ± 0.010
	MRML	0.840 ± 0.035	0.819 ± 0.026	0.874 ± 0.027	0.846 ± 0.022	0.865 ± 0.026	0.807 ± 0.027	0.835 ± 0.022
	LogicDM	0.873 ± 0.033	0.862 ± 0.028	0.850 ± 0.030	0.856 ± 0.034	0.874 ± 0.030	0.850 ± 0.030	0.862 ± 0.034
	BMR	0.901 ± 0.008	0.890 ± 0.014	0.910 ± 0.017	0.891 ± 0.018	0.910 ± 0.012	0.890 ± 0.018	0.891 ± 0.021
	QMFND	0.942 ± 0.011	0.930 ± 0.024	0.950 ± 0.027	0.940 ± 0.022	0.950 ± 0.023	0.930 ± 0.013	0.940 ± 0.022
	MTTV	0.918 ± 0.018	0.893 ± 0.010	0.934 ± 0.023	0.932 ± 0.018	0.936 ± 0.023	0.926 ± 0.013	0.934 ± 0.023
	BMLHF	**0.950** ± 0.012	0.945 ± 0.016	0.955 ± 0.010	**0.950** ± 0.017	0.955 ± 0.017	0.945 ± 0.020	**0.950** ± 0.018

4.3.2. Overall performance

Table 1 shows the fake news detection results on four measurements, including Accuracy, Precision, Recall and F1 scores, respectively. We obtain the following observations:

Specifically, our approach achieves an accuracy improvement of 0.015 (=0.912–0.897) on Weibo, 0.048 (=0.966–0.918) on Twitter, and 0.008 (=0.950–0.942) on Fakeddit. For F1 of fake news, BMLHF achieves an improvement of 0.002 (=0.903–0.901) on Weibo, 0.022 (=0.957–0.935) on the Twitter dataset, and 0.010 (=0.950–0.940) on Fakeddit. For F1 of real news, BMLHF achieves an improvement of 0.010 (=0.902–0.892) on Weibo, 0.046 (=0.956–0.910) on Twitter, and 0.010 (=0.950–0.940) on Fakeddit. From the table, some compared methods can obtain slightly better results on specific metrics. However, they have relatively weaker results on the other metrics, including the comprehensive metrics than BMLHF. These results demonstrate that BMLHF effectively considers and alleviates the issue of modality imbalance. Furthermore, BMLHF learns multi-modal feature representations with stronger discriminability and better captures multi-view information within and between modalities.

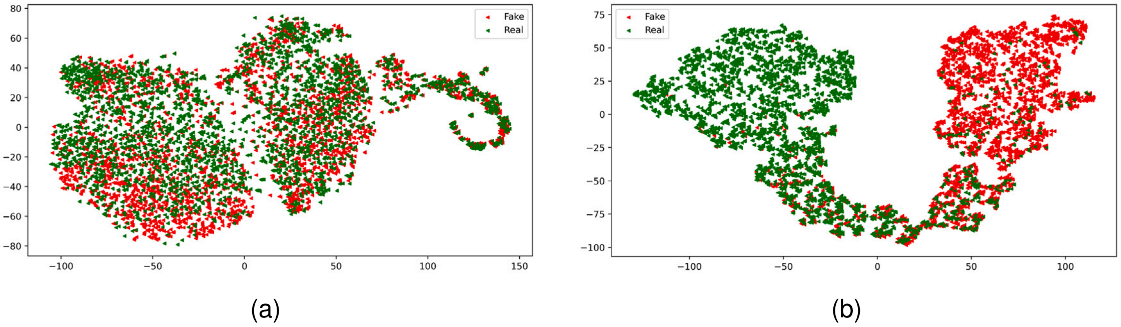
Figs. 4, 5, and 6 show the T-SNE diagrams of sample distribution on Weibo, Twitter, and Fakeddit respectively. For original samples, we simply concatenate the original features of the text and image. In Figs. 4(a), 5(a), and 6(a) we can see that real and fake news mix together, whereas in Figs. 4(b), 5(b), and 6(b) we can find that real and fake news exhibit relatively good separability, with few samples are not completely separated. This demonstrates the effectiveness of BMLHF for multi-modal fake news classification.

4.4. Discussions

4.4.1. Evaluation of key modules

In order to evaluate the influence of key modules of BMLHF, we conduct ablation experiments. Table 2 shows the results of ablation experiments. We remove **MIB**, which we call **BMLHF-I**. We remove the two-stage fusion block of **HF**, which refers to **BMLHF-T**, and we directly concatenate the text and image features after the dual cross-transformer interaction block. We remove the dual cross-transformer block of **HF** and call it **BMLHF-D**, and multi-view features are fed into the two-stage fusion block without interaction. We remove the whole **HF** including the dual cross-transformer block and the two-stage fusion block, which we call **BMLHF-H**. For **BMLHF-H**, we directly concatenate the text and image features. We directly concatenate the intra-modal features without using our first fusion stage of **HF**, which we call **BMLHF-tra**, and we directly concatenate the inter-modal features without using our second fusion stage of **HF**, which we call **BMLHF-ter**.

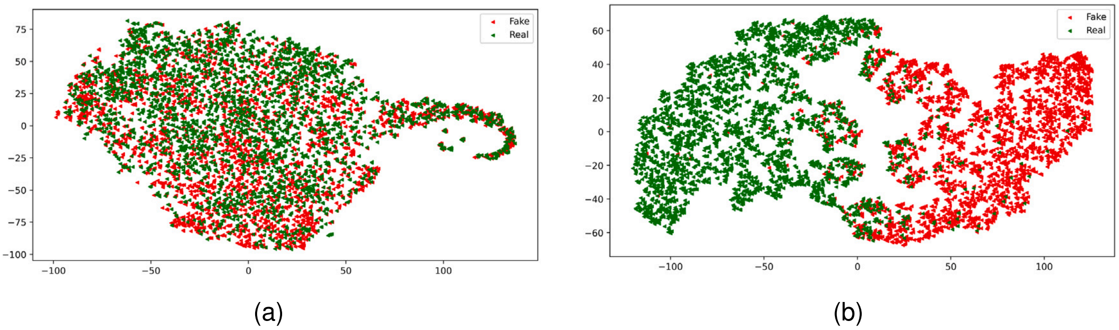
For **BMLHF-I**, there is a significant performance reduction compared to complete model, with a decrease of 0.038 (=0.912–0.874) in accuracy on Weibo and 0.040 (=0.966–0.926) on Twitter. This highlights the advantage of **MIB** in balancing multi-modal information. For **BMLHF-T**, the accuracy declines by 0.022 (=0.912–0.890) on Weibo and 0.024 (=0.966–0.942) on Twitter, demonstrating the positive effect of the two-stage fusion strategy of **HF** in effectively integrating multi-view features and exploring cross-modal correlation within and between modalities. For **BMLHF-D**, there is a drop of 0.039 (=0.912–0.873) and 0.041 (=0.966–0.925) in accuracy on Weibo and Twitter, respectively, indicating the importance of the dual cross-transformer block of **HF** in facilitating interaction between modalities.



(a)

(b)

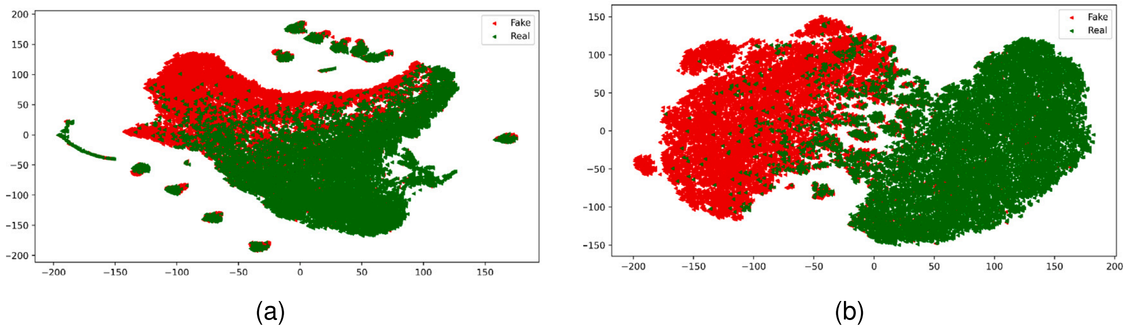
Fig. 4. T-SNE of sample distribution on Weibo, where (a) shows the distribution of original samples, and (b) shows the distribution of learned features.



(a)

(b)

Fig. 5. T-SNE of sample distribution on Twitter, where (a) shows the distribution of original samples, and (b) shows the distribution of learned features.



(a)

(b)

Fig. 6. T-SNE of sample distribution on Fakeddit, where (a) shows the distribution of original samples, and (b) shows the distribution of learned features.

For **BMLHF-H**, there is a significant performance reduction compared to complete model, with a decrease of 0.044 (=0.912−0.868) in accuracy on the Weibo, 0.056 (=0.966−0.910) on the Twitter, and 0.030 (0.950−0.920) on the Fakeddit. This highlights the advantage of HF in effectively interacting and integrating features. For **BMLHF-tra**, the accuracy declines by 0.010 (=0.912−0.902) on Weibo, 0.033 (=0.966−0.933) on Twitter, and 0.015 (=0.950−0.935) on Fakeddit. This highlights the advantage of the first fusion stage of HF in sufficiently fusing semantic and pattern features within modalities. For **BMLHF-ter**, there is a drop of 0.019 (=0.912−0.893), 0.036 (=0.966−0.930), 0.014 (0.950−0.936) in accuracy on Weibo, Twitter and Fakeddit respectively. This highlights the advantage of the second fusion stage of HF in effectively obtaining fused features between modalities. Overall, the results shown in Table 2 illustrate the effectiveness of each module of BMLHF.

4.4.2. Evaluation of imbalanced multi-modal learning

To further verify the validity of MIB, we experimentally observe the variation of modal information on three datasets. According to Eq. (2), s_i^u quantifies the information of different modalities, and we regard s represents the average information amount of a batch of samples. We observe the value of s to reflect whether MIB can help balance the information of multiple modalities. Fig. 7 shows the experimental results on three datasets.

We can observe that the text modality contains more information than the image modality on Weibo, Twitter and Fakeddit. Moreover, as the training progresses, the difference in information between the modalities is decreased. Results on the Weibo, Twitter and Fakeddit datasets indicate that the difference of information of text and image is effectively decreased during the training process, demonstrating the effectiveness of the MIB module in alleviating the issue of modality imbalance.

Furthermore, we apply our designed MIB module to these methods mentioned in Section 1.1 for helping balance information of different modalities. Specifically, we have set up dimensionality reduction networks after each feature extractor. MIB will monitor modal information and assign corresponding gradient update weights of each modality. The experimental results are shown in Fig. 8. From Fig. 8, we can find that the results of the complete multi-modal version with MIB show a significant improvement on accuracy as compared with the original multi-modal version for these three methods, which fully verifies MIB can effectively alleviate modality imbalance problem and improve the performance of model.

Table 2
Ablation studies of BMLHF on Weibo, Twitter, and Fakeddit datasets.

Dataset	Method	Acc	Fake news			Real news		
			Pre	Rec	F1	Pre	Rec	F1
Weibo	BMLHF-I	0.874	0.930	0.830	0.870	0.830	0.930	0.870
	BMLHF-T	0.890	0.900	0.885	0.890	0.885	0.900	0.890
	BMLHF-D	0.873	0.890	0.860	0.880	0.850	0.880	0.870
	BMLHF-H	0.868	0.880	0.855	0.866	0.853	0.880	0.866
	BMLHF-tra	0.902	0.890	0.900	0.895	0.860	0.880	0.875
	BMLHF-ter	0.893	0.870	0.900	0.885	0.900	0.870	0.885
	BMLHF	**0.912**	0.930	0.880	**0.903**	0.894	0.920	**0.902**
Twitter	BMLHF-I	0.926	0.986	0.880	0.922	0.880	0.986	0.922
	BMLHF-T	0.942	0.950	0.935	0.945	0.935	0.950	0.945
	BMLHF-D	0.925	0.943	0.912	0.933	0.901	0.933	0.922
	BMLHF-H	0.910	0.899	0.910	0.903	0.920	0.899	0.910
	BMLHF-tra	0.933	0.915	0.913	0.914	0.935	0.915	0.920
	BMLHF-ter	0.930	0.920	0.930	0.925	0.930	0.920	0.925
	BMLHF	**0.966**	0.985	0.933	**0.957**	0.948	0.975	**0.956**
Fakeddit	BMLHF-I	0.905	0.880	0.940	0.910	0.940	0.880	0.915
	BMLHF-T	0.923	0.895	0.945	0.930	0.940	0.900	0.925
	BMLHF-D	0.937	0.920	0.950	0.940	0.950	0.920	0.940
	BMLHF-H	0.920	0.890	0.945	0.925	0.945	0.890	0.925
	BMLHF-tra	0.935	0.940	0.930	0.935	0.930	0.940	0.935
	BMLHF-ter	0.936	0.935	0.925	0.930	0.925	0.935	0.930
	BMLHF	**0.950**	0.945	0.955	**0.950**	0.955	0.945	**0.950**

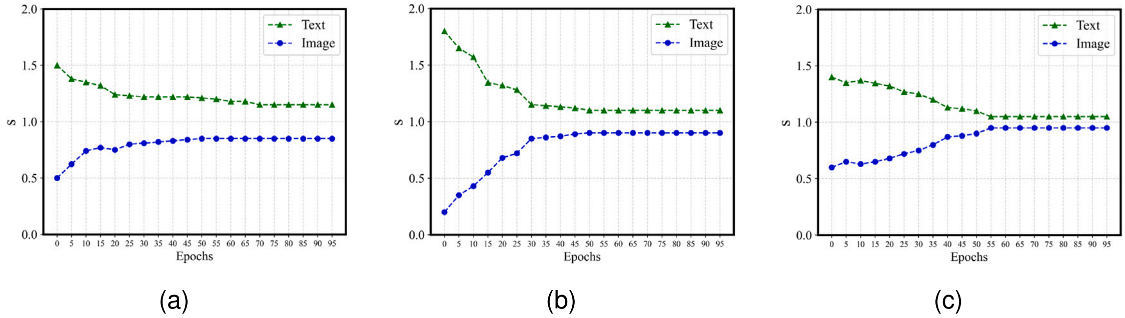


Fig. 7. The variation curve of information of modalities on (a) Weibo, (b) Twitter, (c) Fakeddit.

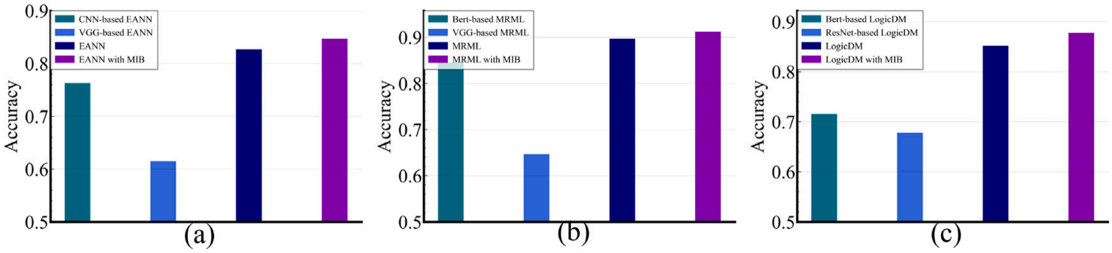


Fig. 8. Overall accuracy between uni-modal versions, multi-modal version, and MIB-based version on Weibo. (a) Text-only EANN, image-only EANN, EANN, and EANN with MIB. (b) Text-only MRML, image-only MRML, MRML, MRML with MIB. (c) Text-only LogicDM, image-only LogicDM, LogicDM, LogicDM with MIB.

4.4.3. Hyperparameters analysis

We discuss the sensitivity of our approach to different values of hyperparameters τ, α_1 and α_2 on Weibo, Twitter, and Fakeddit. Figs. 9, 10 and 11 show the experimental results of evaluation of τ, α_1 and α_2. From the figures, for τ, BMLHF performs better in the range [0.3,0.5] on Weibo, 0.5 on Twitter and Fakeddit. For α_1, BMLHF performs better in the range [0.7,0.9] on Weibo and at 0.8 on Twitter and Fakeddit. For α_2, BMLHF performs better in the range [0.4,0.6] on Weibo and at 0.5 on Twitter and Fakeddit. Therefore, we set τ=0.5, α_1=0.8, α_2=0.5 for three datasets.

Furthermore, we have fixed the optimal hyperparameters to observe the variation of performance during different epochs. Specifically, we adopt the optimal hyperparameters $\tau = 0.5$, $\alpha_1 = 0.8$, $\alpha_2 = 0.5$ on Weibo, Twitter and Fakeddit. Fig. 12 shows the results. We can find that with the fixed optimal hyperparameters, BMLHF tends to achieve the best result at about 70 epochs on three datasets. Then BMLHF can obtain stable performance with increasing epochs.

4.4.4. Statistical analysis

To statistically analyze the difference between our approach and compared methods, we conduct the Kruskal Wallis test [53] at significance level of 0.05 to perform statistical significance test between BMLHF and other methods, Table 3 shows the P-value at significance level of 0.05 between BMLHF and other methods. When P-value is lower than 0.05, we can consider that BMLHF has a significant difference against the corresponding comparison methods. From Table 3, we

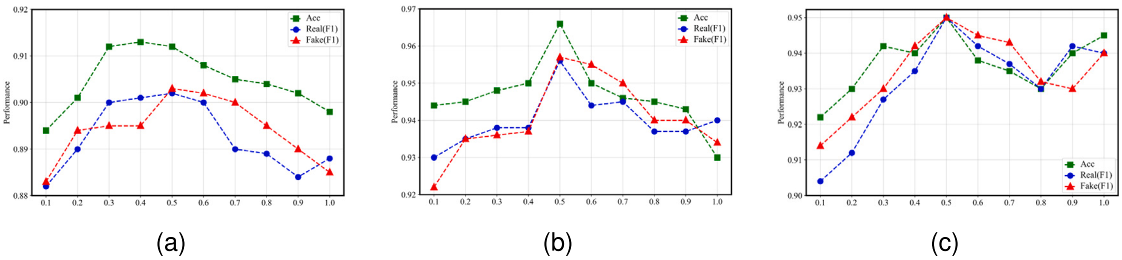


Fig. 9. The performances of BMLHF with different values of τ on (a) Weibo, (b) Twitter, (c) Fakeddit.

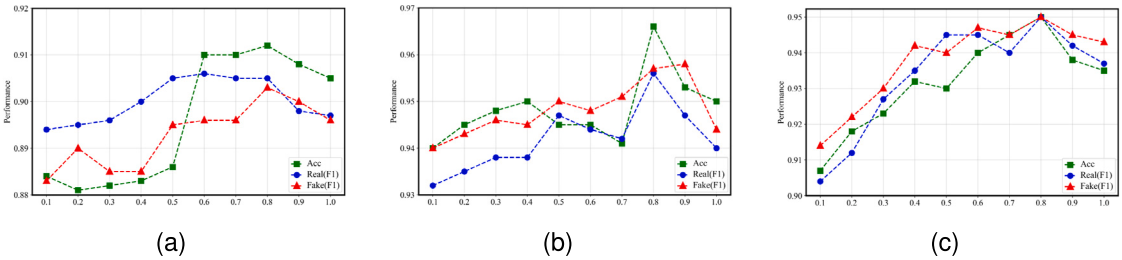


Fig. 10. The performances of BMLHF with different values of α_1 on (a) Weibo, (b) Twitter, (c) Fakeddit.

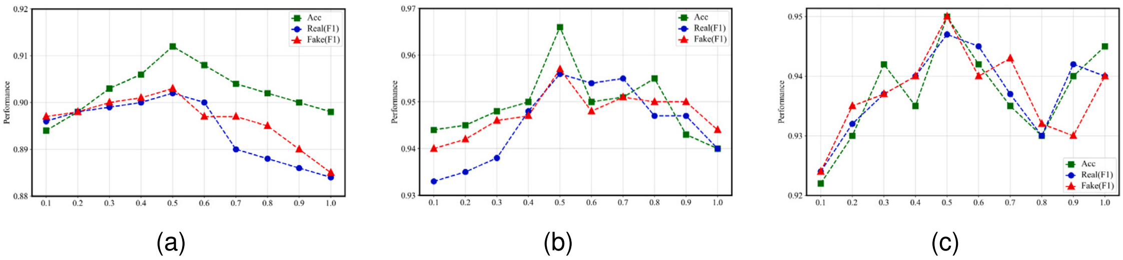


Fig. 11. The performances of BMLHF with different values of α_2 on (a) Weibo, (b) Twitter, (c) Fakeddit.

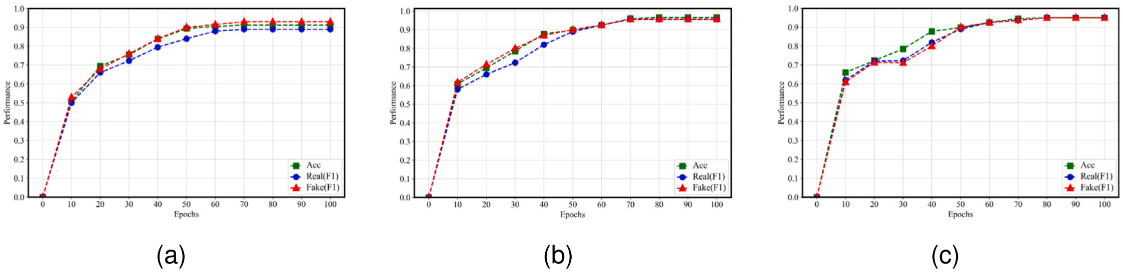


Fig. 12. The performances of BMLHF with fixed hyperparameters on (a) Weibo, (b) Twitter, (c) Fakeddit.

can find the P-values are significantly lower than 0.05. This demonstrates BMLHF indeed has a significant difference from the comparison methods.

4.4.5. Computational cost analysis

In order to investigate computational cost of training and testing phases, we calculate the training time (for each epoch) and testing time (on the total test set) on Weibo, Twitter and Fakeddit. In addition, we also calculate the number of parameters of our approach. The detailed computational cost and number of parameters of our approach and compared methods that can achieve favorable results, i.e., MRML and LogicDM, are reported in Table 4. From the table, training time, testing time, and number of parameters of BMLHF is comparable to MRML and LogicDM. Overall, the computational cost of BMLHF is mainly from the cross-modal transformer part. This comparison indicates that our approach does not need extra large amount of cost, and it is computationally efficient.

5. Conclusion

In this paper, we propose an approach called Balanced Multi-modal Learning with Hierarchical Fusion (BMLHF) for MFND. Specifically, we design a Multi-modal Information Balancing (MIB) module, which

Table 3
P-values between BMLHF and other compared methods on three datasets.

Dataset	BMLHF					
	BERT	Swin-T	EANN	SpotFake+	SAFE	CAFE
Weibo	$2.355 \times e^{-5}$	$1.833 \times e^{-10}$	$9.888 \times e^{-5}$	$3.000 \times e^{-3}$	$5.111 \times e^{-7}$	$1.351 \times e^{-6}$
Twitter	$1.941 \times e^{-6}$	$2.563 \times e^{-9}$	$6.149 \times e^{-6}$	$6.332 \times e^{-9}$	$5.625 \times e^{-10}$	$2.878 \times e^{-11}$
Fakeddit	$1.377 \times e^{-6}$	$1.612 \times e^{-5}$	$1.537 \times e^{-6}$	$1.102 \times e^{-9}$	$4.109 \times e^{-8}$	$2.000 \times e^{-3}$
	MRML	LogicDM	BMR	QMFND	MTTV	
Weibo	$9.226 \times e^{-6}$	$4.187 \times e^{-6}$	$7.505 \times e^{-8}$	$2.600 \times e^{-2}$	$4.000 \times e^{-3}$	
Twitter	$2.052 \times e^{-7}$	$8.000 \times e^{-3}$	$1.000 \times e^{-3}$	$2.800 \times e^{-2}$	$3.692 \times e^{-6}$	
Fakeddit	$4.778 \times e^{-8}$	$4.366 \times e^{-11}$	$1.158 \times e^{-8}$	$1.500 \times e^{-2}$	$1.000 \times e^{-3}$	

Table 4
Computational cost of the training and testing phases of BMLHF on three datasets.

Dataset	Training time (s)			Testing time (s)			Parameters (M)
	Weibo	Twitter	Fakeddit	Weibo	Twitter	Fakeddit	
MRML	37.28	40.23	56.00	24.50	21.98	35.57	13.55
LogicDM	53.80	51.25	85.38	42.44	39.30	56.20	18.05
BMLHF	41.66	38.40	63.50	25.06	22.50	34.70	14.85

accelerates balance among diverse modal information during the optimization process, with corresponding weights assigned to different modalities to inhibit the optimization of the dominate modality. We design a Hierarchical Fusion (HF) module from the within-modality and between-modality fusion perspectives, which effectively leverages the multi-view information and fully explores the correlation within and between modalities.

Comprehensive experiments on two widely used datasets demonstrate the validity of BMLHF. Our model improves the accuracy of fake news detection and outperforms state-of-the-art MFND methods. The ablation studies show the superiority of MIB in alleviating modality imbalance and HF in better fusing multi-view and inter-modal complementary information.

Currently, our approach mainly focuses on image and text modalities. In future, BMLHF will be further evaluated on more real-world experimental data with diverse modalities, e.g., video, to demonstrate its generalization, which helps determine the reliability of the model in more complex practical application scenarios.

CRediT authorship contribution statement

Fei Wu: Writing – review & editing, Writing – original draft, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Shu Chen:** Writing – original draft, Validation, Methodology, Investigation. **Guangwei Gao:** Writing – review & editing, Software, Methodology, Investigation, Formal analysis. **Yimu Ji:** Validation, Resources, Funding acquisition, Formal analysis. **Xiao-Yuan Jing:** Supervision, Investigation, Formal analysis, Conceptualization.

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 was supported by National Natural Science Foundation of China (No. 62076139), National Key Laboratory of Information System Engineering (No. 05202305), and the Open Research Fund of The State Key Laboratory of Blockchain and Data Security, Zhejiang University.

Data availability

Data will be made available on request.

References

[1] C. Boididou, S. Papadopoulos, M. Zampoglou, L. Apostolidis, O. Papadopoulou, Y. Kompatsiaris, Detection and visualization of misleading content on Twitter, Int. J. Multimed. Inf. Retr. (2018) 71–86.

[2] Z. Jin, J. Cao, Y. Zhang, J. Luo, News verification by exploiting conflicting social viewpoints in microblogs, in: AAAI Conference on Artificial Intelligence, 2016, pp. 2972–2978.

[3] Z. Jin, J. Cao, H. Guo, Y. Zhang, J. Luo, Multimodal fusion with recurrent neural networks for rumor detection on microblogs, in: ACM International Conference on Multimedia, 2017, pp. 795–816.

[4] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science 359 (6380) (2018) 1146–1151.

[5] F. Olan, U. Jayawickrama, E.O. Arakpogun, J. Suklan, S. Liu, Fake news on social media: the impact on society, Inf. Syst. Front. 26 (2) (2024) 443–458.

[6] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, L. Wei, Detecting fake news by exploring the consistency of multimodal data, Inf. Process. Manage. (2021) 102610.

[7] L. Peng, S. Jian, Z. Kan, L. Qiao, D. Li, Not all fake news is semantically similar: Contextual semantic representation learning for multimodal fake news detection, Inf. Process. Manage. 61 (1) (2024) 103564.

[8] M. Mousoulidou, L. Taxitari, A. Christodoulou, Social media news headlines and their influence on well-being: Emotional states, emotion regulation, and resilience, Eur. J. Investig. Heal. Psychol. Educ. 14 (6) (2024) 1647–1665.

[9] Q. Liao, H. Chai, H. Han, X. Zhang, X. Wang, W. Xia, Y. Ding, An integrated multi-task model for fake news detection, IEEE Trans. Knowl. Data Eng. 34 (11) (2021) 5154–5165.

[10] X. Zhou, K. Shu, V.V. Phoha, H. Liu, R. Zafarani, "This is fake! shared it by mistake": Assessing the intent of fake news spreaders, in: ACM Web Conference, 2022, pp. 3685–3694.

[11] A. Mosallanezhad, M. Karami, K. Shu, M.V. Mancenido, H. Liu, Domain adaptive fake news detection via reinforcement learning, in: ACM Web Conference, 2022, pp. 3632–3640.

[12] Q. Nan, J. Cao, Y. Zhu, Y. Wang, J. Li, MDFEND: Multi-domain fake news detection, in: Association for Computing Machinery, 2021, pp. 3343–3347.

[13] B. Yu, W. Li, X. Li, J. Zhou, J. Lu, Uncertainty-aware hierarchical labeling for face forgery detection, Pattern Recognit. 153 (2024) 110526.

[14] B. Hu, Q. Sheng, J. Cao, Y. Shi, Y. Li, D. Wang, P. Qi, Bad actor, good advisor: Exploring the role of large language models in fake news detection, in: AAAI Conference on Artificial Intelligence, 2024, pp. 22105–22113.

[15] A. Sharif Razavian, H. Azizpour, J. Sullivan, S. Carlsson, CNN features off-the-shelf: An astounding baseline for recognition, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, pp. 806–813.

[16] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2014.

[17] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, J. Gao, Eann: Event adversarial neural networks for multi-modal fake news detection, in: ACM Sigkdd International Conference on Knowledge Discovery & Data Mining, 2018, pp. 849–857.

[18] S. Singhal, A. Kabra, M. Sharma, R.R. Shah, T. Chakraborty, P. Kumaraguru, Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract), in: AAAI Conference on Artificial Intelligence, 2020, pp. 13915–13916.

[19] X. Zhou, J. Wu, R. Zafarani, Similarity-aware multi-modal fake news detection, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2020, pp. 354–367.

[20] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, L. Shang, Cross-modal ambiguity learning for multimodal fake news detection, in: ACM Web Conference, 2022, pp. 2897–2905.

[21] L. Peng, S. Jian, D. Li, S. Shen, MRML: Multimodal rumor detection by deep metric learning, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2023, pp. 1–5.

[22] H. Liu, W. Wang, H. Li, Interpretable multimodal misinformation detection with logic reasoning, Assoc. Comput. Linguist. (2023) 9781–9796.

[23] Q. Ying, X. Hu, Y. Zhou, Z. Qian, D. Zeng, S. Ge, Bootstrapping multi-view representations for fake news detection, in: AAAI Conference on Artificial Intelligence, 2023, pp. 5384–5392.

[24] T. Winterbottom, S. Xiao, A. McLean, N.A. Moubayed, On modality bias in the tvqa dataset, in: British Machine Vision Conference, 2020, pp. 1–20.

[25] C. Du, T. Li, Y. Liu, Z. Wen, T. Hua, Y. Wang, H. Zhao, Improving multi-modal learning with uni-modal teachers, 2021, arXiv preprint arXiv:2106.11059.

[26] X. Peng, Y. Wei, A. Deng, D. Wang, D. Hu, Balanced multimodal learning via on-the-fly gradient modulation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8238–8247.

[27] Y. Wei, R. Feng, Z. Wang, D. Hu, Enhancing multimodal cooperation via sample-level modality valuation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 27338–27347.

[28] R. Xu, R. Feng, S.-X. Zhang, D. Hu, Mmcosine: Multi-modal cosine loss towards balanced audio-visual fine-grained learning, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2023, pp. 1–5.

[29] K. Nakamura, S. Levy, W.Y. Wang, r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection, 2019, arXiv preprint arXiv:1911.03854.

[30] X. Zhou, R. Zafarani, A survey of fake news: Fundamental theories, detection methods, and opportunities, ACM Comput. Surv. (2020) 1–40.

[31] X. Zhang, A.A. Ghorbani, An overview of online fake news: Characterization, detection, and discussion, Inf. Process. Manage. (2020) 102025.

[32] X. Zhou, R. Zafarani, K. Shu, H. Liu, Fake news: Fundamental theories, detection strategies and challenges, in: ACM International Conference on Web Search and Data Mining, 2019, pp. 836–837.

[33] J. Ma, W. Gao, P. Mitra, S. Kwon, B.J. Jansen, K.-F. Wong, M. Cha, Detecting rumors from microblogs with recurrent neural networks, in: International Joint Conference on Artificial Intelligence, 2016, pp. 3818–3824.

[34] V. Vaibhav, R.M. Annasamy, E. Hovy, Do sentence interactions matter? Leveraging sentence level representations for fake news classification, in: Conference on Empirical Methods in Natural Language Processing, 2019, pp. 134–139.

[35] A. Giachanou, P. Rosso, F. Crestani, Leveraging emotional signals for credibility detection, in: International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 877–880.

[36] K. Wu, S. Yang, K.Q. Zhu, False rumors detection on sina weibo by propagation structures, in: IEEE International Conference on Data Engineering, 2015, pp. 651–662.

[37] F. Yang, Y. Liu, X. Yu, M. Yang, Automatic detection of rumor on sina weibo, in: ACM SIGKDD Workshop on Mining Data Semantics, 2012, pp. 1–7.

[38] Z. Jin, J. Cao, Y. Zhang, J. Zhou, Q. Tian, Novel visual and statistical image features for microblogs news verification, IEEE Trans. Multimed. (3) (2016) 598–608.

[39] C. Boididou, S. Papadopoulos, D.-T. Dang-Nguyen, G. Boato, Y. Kompatsiaris, et al., The CERTH-UNITN participation@ verifying multimedia use 2015, MediaEval 1 (2015) 2.

[40] D. Khattar, J.S. Goud, M. Gupta, V. Varma, Mvae: Multimodal variational autoencoder for fake news detection, in: The World Wide Web Conference, 2019, pp. 2915–2921.

[41] P. Qi, J. Cao, T. Yang, J. Guo, J. Li, Exploiting multi-domain visual information for fake news detection, in: IEEE International Conference on Data Mining, 2019, pp. 518–527.

[42] Y. Luo, J. Ma, C.K. Yeo, BCMM: A novel post-based augmentation representation for early rumour detection on social media, Pattern Recognit. 113 (2021) 107818.

[43] A. Bondielli, P. Dell'Oglio, A. Lenci, F. Marcelloni, L.C. Passaro, M. Sabbatini, Multi-fake-detective at evalita 2023: Overview of the multimodal fake news detection and verification task, in: Evaluation Campaign of Natural Language Processing and Speech Tools, 2023.

[44] S. Tufchi, A. Yadav, T. Ahmed, A comprehensive survey of multimodal fake news detection techniques: Advances, challenges, and opportunities, Int. J. Multimed. Inf. Retr. 12 (2) (2023) 28.

[45] S. Qian, J. Wang, J. Hu, Q. Fang, C. Xu, Hierarchical multi-modal contextual attention network for fake news detection, in: International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 153–162.

[46] Z. Qu, Y. Meng, G. Muhammad, P. Tiwari, QMFND: A quantum multimodal fusion-based fake news detection model for social media, Inf. Fusion 104 (2024) 102172.

[47] W. Wang, D. Tran, M. Feiszli, What makes training multi-modal classification networks hard? in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12695–12705.

[48] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.

[49] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[50] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, I. Gretchen, Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, 2021, pp. 8748–8763.

[51] S. Liu, X. Yue, F. Wu, J. Sun, Y. Feng, Y. Ji, Semantic distillation and structural alignment network for fake news detection, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2024, pp. 6620–6624.

[52] B. Wang, Y. Feng, X.-c. Xiong, Y.-h. Wang, B.-h. Qiang, Multi-modal transformer using two-level visual features for fake news detection, Appl. Intell. 53 (9) (2023) 10429–10443.

[53] Y. Chan, R.P. Walmsley, Learning and understanding the Kruskal-Wallis one-way analysis-of-variance-by-ranks test for differences among three or more independent groups, Phys. Ther. 77 (12) (1997) 1755–1761.

Fei Wu received the Ph.D. degree in Information and Communication Engineering from Nanjing University of Posts and Telecommunications (NJUPT), China, in 2016. He is currently a professor with the College of Automation and College of Artificial Intelligence in NJUPT. He has authored over eighty scientific papers, such as TPAMI, TIP, TCYB, PR, TSE, TR, CVPR, AAAI, IJCAI and WWW. His research interests include pattern recognition and social media analysis.

Shu Chen is pursuing the Master degree in electronic information from Nanjing University of Posts and Telecommunications, Nanjing, China. His research interests include multi-modal learning.

Guangwei Gao received the Ph.D. degree in pattern recognition and intelligence systems from Nanjing University of Science and Technology, Nanjing, China, in 2014. Now, he is an associate professor with the Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing, China. His research mainly focuses on pattern recognition and computer vision.

Yimu Ji is a professor in Nanjing University of Posts and Telecommunications, Nanjing, China. His research mainly focuses on intelligent information processing.

Xiao-Yuan Jing received the doctoral degree of pattern recognition and intelligent system in the Nanjing University of Science and Technology, 1998. Now he is a professor with the School of Computer, Wuhan University, China. His research interests include artificial intelligence and pattern recognition.