Multi-modal Fake News Detection on Social Media via Multi-grained Information Fusion

Yangming Zhou
ymzhou21@m.fudan.edu.cn
School of Computer Science
Fudan University
Shanghai, China

Yuzhou Yang
yzyang22@m.fudan.edu.cn
School of Computer Science
Fudan University
Shanghai, China

Qichao Ying
shinydotcom@163.com
School of Computer Science
Fudan University
Shanghai, China

Zhenxing Qian*
zxqian@fudan.edu.cn
School of Computer Science
Fudan University
Shanghai, China

Xinpeng Zhang*
zhangxinpeng@fudan.edu.cn
School of Computer Science
Fudan University
Shanghai, China

ABSTRACT

The easy sharing of multimedia content on social media has caused a rapid dissemination of fake news, which threatens society's stability and security. Therefore, fake news detection has garnered extensive research interest in the field of social forensics. Current methods primarily concentrate on the integration of textual and visual features but fail to effectively exploit multi-modal information at both fine-grained and coarse-grained levels. Furthermore, they suffer from an ambiguity problem due to a lack of correlation between modalities or a contradiction between the decisions made by each modality. To overcome these challenges, we present a Multi-grained Multi-modal Fusion Network (MMFN) for fake news detection. Inspired by the multi-grained process of human assessment of news authenticity, we respectively employ two Transformer-based pre-trained models to encode token-level features from text and images. The multi-modal module fuses fine-grained features, taking into account coarse-grained features encoded by the CLIP encoder. To address the ambiguity problem, we design uni-modal branches with similarity-based weighting to adaptively adjust the use of multi-modal features. Experimental results demonstrate that the proposed framework outperforms state-of-the-art methods on three prevalent datasets.

CCS CONCEPTS

• **Information systems → Data mining**.

KEYWORDS

Fake news detection, Multi-modal learning, Multi-modal fusion

ACM Reference Format:
Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. 2018. Multi-modal Fake News Detection on Social Media via Multi-grained Information Fusion. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (ICMR '23).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

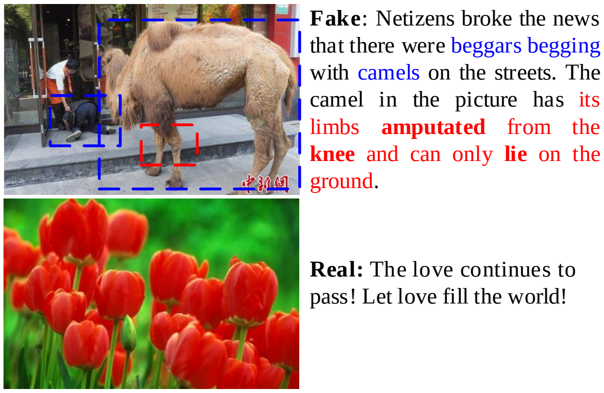
Online social networks, such as Twitter and Weibo, have largely replaced traditional forms of information communication represented by newspapers and magazines. Despite the convenience they offer to users, including the ability to seek out friends and share viewpoints, these networks have also facilitated the rapid and widespread dissemination of fake news [16, 25, 47]. Compared to traditional written materials, online news posts are more susceptible to manipulation. Moreover, readers can be easily influenced by well-crafted fake news and may inadvertently further its spread.The consequences of fake news include the creation of panic, the misdirection of public opinion, and negative impacts on society. With hundreds of millions of social media users generating a vast amount of posts constantly, the use of traditional, inefficient manual review methods is not feasible for mitigating the threat of fake news. In response, many researchers in the field of computer science have recently focused on developing methods for detecting fake news [2, 7, 21, 24, 29].

Early works on fake news detection primarily focused on analyzing either text-only or image-only content [5, 9]. These works typically verify the logical and semantic coherence of the input and took into account trivial indicators, such as grammatical errors or traces of image manipulation. Although uni-modal approaches are effective, modern news articles and posts often include multiple modalities that are inter-related, hence, these methods overlook such correlations. For instance, a real image can be combined with total rumors and correct words can be used to describe a tampered image. Given this, multi-modal feature analysis is necessary to offer complementary benefits for fake news detection. In recent years, there have been many works that combine multi-modal features to detect anomalies in news [7, 16, 35, 41]. Despite the advancements made in the field, current approaches face two significant challenges. First, though many works come up with novel fusion methods, they

[*] Corresponding author. All the authors are also with the Key Laboratory of Culture and Tourism Intelligent Computing of Ministry of Culture and Tourism, Fudan University.

ICMR '23, June 12-15, 2023, Thessaloniki, Greece

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX



Fake: Netizens broke the news that there were beggars begging with camels on the streets. The camel in the picture has its limbs **amputated** from the **knee** and can only **lie** on the ground.

Real: The love continues to pass! Let love fill the world!

Figure 1: Examples of the two challenges faced by current works. Upper: multi-grained information is required to detect misinformation (the blue box represents matched elements and the red box represents mismatched elements). Lower: weak correlation between text and image may prevent users/models from finding cross-modal clues.

merely fuse them at holistic post level, inevitably missing some detailed information; Or they only consider the matching between entities, tokens, or regions, neglecting global semantic correlation. This results in suboptimal utilization of information across different granularities. Second, many works rely excessively on multi-modal fusion features, thus suffer from the ambiguity problem, i.e. inconsistency caused by inter-modal conflict or weak correlation.

Figure 1 illustrates two examples in the Weibo dataset that demonstrate the aforementioned two challenges. The upper image in Figure 1 depicts the process of multi-grained fake news detection, wherein neither uni-modal feature of the text nor image is capable of verifying authenticity. First, readers of the post would pay attention to the objects in the image and entities in the text. For the examples given, people would first see the beggar and camel in the picture, and the words *beggar*, *begging*, *camel*, *amputation*, *knee* and *lie* in the text. At this time, other than matched elements (marked as blue region), they would find the three words *amputation*, *knee* and *lie* do not match the image content (marked as red region). Subsequently, they will comprehend the semantic meaning of the sentence and image as a whole, perform an analysis to determine whether the two match, and finally, arrive at a conclusion regarding the authenticity of the news. Many existing works tend to ignore this point. The lower image of Figure 1 is an example that illustrates the ambiguity and is also a common type of post on social media. The visual objects and the textual entities in the post have no matching relationship and are semantically unrelated. The user who posts this post may only be expressing their feelings. Manual review can easily determine that this is not a fake news and is not harmful. However, models that over-emphasize multi-modal fusion may misjudge it due to the mismatches in the multi-modal features. As a result, the post may not pass the automatic review process of social software and may not be published smoothly. At this point, analyzing uni-modal content and emotions is sufficient to evaluate the credibility of the post, and multi-modal fusion features are not necessary and may even add noise to the classification task. To address this issue, Chen et al. [7] proposed CAFE that first compresses the image and text and contrasts the news that has the correct image-text pairs to learn by minimizing the Kullback-Leibler (KL) divergence. Then, the corresponding cross-modal ambiguity score is used to reweight the multimodal features. However, as only limited data is input into the network, their alignment of multi-modal features cannot be guaranteed. Furthermore, they process uni-modal features through variational autoencoders with non-shared weights, so we cannot determine if the KL divergence is reliable enough to measure cross-modal ambiguity. Thus, resolving the ambiguity problem remains a motivation for our research work.

To address the above-mentioned issues, this paper proposes the Multi-grained Multi-modal Fusion Network (MMFN).The MMFN approach integrates uni-modal features and a multi-grained multi-modal fused feature for more accurate fake news detection. MMFN encodes text and image using pre-trained BERT [11] and Swin Transformer (Swin-T) [19] models respectively, which extract fine-grained information at the token level. The pre-trained CLIP [27] model is utilized to encode coarse-grained features and capture semantic information at the post level, which is further used to address the ambiguity problem. We propose a multi-grained multi-modal fusion module to perform granularity-level fusion on multi-modal features, in which fine-grained features are fed into co-attention Transformer (CT) blocks for generating aligned fine-grained multi-modal features, coarse-grained features are used for evaluating the cross-modal correlation. This module adjust the usage of multi-modal features dynamically, mitigating the ambiguity problem.

The contributions of this paper are mainly three-folded, namely:

- We propose MMFN, which implements the idea of processing multi-modal features at different levels of granularity to form a comprehensive representation that reflects both the detailed and global aspects of the news.
- We specifically design two uni-modal branches and adopt CLIP pre-trained model to evaluate cross-modal correlation, further address the problem brought by scenario with high cross-modal ambiguity.
- We conduct comprehensive experiments on three famous datasets, where MMFN outperforms state-of-the-art fake news detection methods. What's more, ablation studies verify the effectiveness of granularity-level processing and multi-modal features adjustment.

2 RELATED WORKS

2.1 Fake News Detection

Modalities refer to information perceived by different propagation channels that human communication is adapted to [3]. Fake news detection methods could be categorized by utilized modal as uni-modal and multi-modal fake news detection, tons of work has been done on both methods based on deep learning in recent years.

While uni-modal features like text content plays a key role in distinguishing fake news [22, 46], correlation and consistency of multi-modal features are also vital. Earlier works design sophisticated yet black-box attention mechanisms for multi-modal feature fusion [4, 5, 16]. Singhal et al. integrate pre-trained XLNet ResNet respectively for text and feature extraction. The features are then concatenated for the final binary classification [30]. Singhal et al. [32] propose the Spotfake using VGG and BERT to extract features, which is then improved in Spotfake+ [30] for full-length article detection. Wang et al. [35] propose an event advisory neural network, namely EANN, which tries to extract shared features among news of different events. Many other works [7, 16, 41] also propose better fuison of extracted features from different modalities before sending them into a classifier.

Compared with uni-modal methods, multi-modal methods not simply have extra image features, but have more potential information underneath multi-modal contents for mining. Some works try to excavate potential information from the datasets. Qi et al. [24] claim former works neglect information that cannot be extracted by a pre-trained extractor, such as entities and texts within images. Thus, they manually extract these information as supplement of text content. Zhang et al. [43] design a novel dual emotion feature descriptor to measure the emotional gap between the publisher and comments and further verify that dual emotion is distinctive between fake and real news. Allein et al. [2] propose DistilBert, which uses latent representations of news articles and user-generated content to guide model learning. Wang et al. [36] propose KMGCN that integrates textual, visual, and knowledge information into a unified framework to model semantic representations and improve accuracy Further, Abdelnabi et al. [1] use online search engine to gather relevant Web evidence and propose Consistency-Checking Network to mimic human reasoning process.

Except from mining information, planning better multi-modal representations' interaction is vital for better detection performance. SAFE [44] calculates the relevance between news textual and visual information. MCAN [38] stacks multiple co-attention layers to fuse the multi-modal features. Qian et al. [26] use BERT to produce hierarchical semantics of text, and use ResNet to produce regional image representations. Then different levels of semantics are fed into co-attention layers with regional image features to achieve hierarchical multi-modal feature fusion. FND-CLIP [45] uses CLIP-exctracting features as multimodal represention and designs a modal-wise attention to aggregat features.

2.2 Pre-trained Model

Vaswani et al. [34] propose Transformer, which quickly become the state-of-the-art method in many NLP tasks. Subsequently, Devlin et al. [11] propose BERT, a large Transformer-based model pre-trained on large corpora. Transformer-based image processing methods also gained merits in their fields. However, a simple application of self-attention to images would require quadratic cost in the number of pixels, causing a problem with input sizes of realistic images. Dosovitskiy et al. [12] directly apply a Transformer architecture on non-overlapping medium-sized image patches for image classification, namely ViT. On top of that, Liu et al. [19] propose Swin-T combining convolution layer with proposed Swin-T layers, achieving linear increase in complexity with image size. Swin-T focuses on general-purpose performance rather than specifically on classification like ViT, yet still achieves the best speed-accuracy trade-off among other methods on image classification.

In the past decade, multi-modal machine learning has received considerable attention in the research community [3]. Neural architectures are employed in tasks that go beyond single modalities, for example, Visual Question Answering (VQA) [10], Visual Commonsense Reasoning (VCR) [40], etc. In these tasks, priors and features from different modalities are required and algorithms or deep networks cannot be effective when provided with only a single modality. Several generic technologies are developed for learning joint representations of image content and natural language. For example, CLIP [27] is a multi-modal model that combines knowledge of language concepts with semantic knowledge of images. Benefiting from the contrastive learning paradigm, the textual and visual features extracted by CLIP can be considered to be aligned in the same semantic space, which can reflect the correlation between the textual content and visual content of news. The CLIP-based Multimodal learning has been used in a lot of downstream tasks [8, 37]. Other multi-modal models, such as Glide [23] and VilBERT [20], have also been utilized for tasks such as text-to-image generation and cross-modal representation learning.

3 METHOD

In this paper, we propose MMFN to improve the accuracy of multimodal fake news detection via multi-grained multi-modal fusion. The network design of MMFN is shown in Figure 2, which consists of multi-modal feature encoder, multi-grained feature fusion, unimodal branches and modality weighting via CLIP similarity, and fake news classifier.

3.1 Multi-modal Feature Encoder

Textual Feature Encoding via BERT. BERT [11] is a popular pre-trained language model built on Transformer that is trained using unsupervised learning on a large corpus and has achieved excellent results in many NLP downstream tasks. Therefore, we use a BERT model to encode features from **T**. The textual content of a news post, which is the concatenation of text and optical character recognition (OCR) extraction from an image, is a sequential list of words denoted as $\mathbf{T} = \left[t_1, t_2, \ldots, t_{n_w}\right]$, where n_w is the number of words. After applying BERT to **T**, the encoded textual feature $\mathbf{T}^b = \left[t_1^b, t_2^b, \ldots, t_{n_w}^b\right]$ is obtained, where $t_i^b \in \mathbb{R}^{d_b}$ is the output for the last hidden state of the $i-$th token in the text embedding and d_b is the dimension of the word embedding.

Visual Feature Encoding via Swin-T. Swin-T [19] is a vision Transformer that produces a hierarchical feature representation based on shifted window self-attention and achieves state-of-the-art performance on many CV tasks such as object detection and semantic segmentation. In this paper, we introduce Swin-T to the task of fake news detection. Given the visual content $\mathbf{V} \in \mathbb{R}^{w \times h}$, Swin-T transforms it to a sequence embeddings $\mathbf{V^s} = \left[v_1^s, v_2^s, \ldots, v_{n_p}^s\right]$, where w and h are the width and height of the image, $v_i^s \in \mathbb{R}^{d_s}$ is the hidden-states at the output of the last layer of the model corresponding to the $i-$th window of the input, n_p is the number of patch in Swin-T, and d_s is the hidden size of the visual embedding.

Multi-modal Feature Encoding via CLIP. CLIP is pre-trained on a massive and diverse dataset, and can embed texts and images into a uniform mathematical space, making it naturally propitious to calculate the cross-modal correlation. Besides, experiments have shown the model's robustness in dealing with distribution shift,

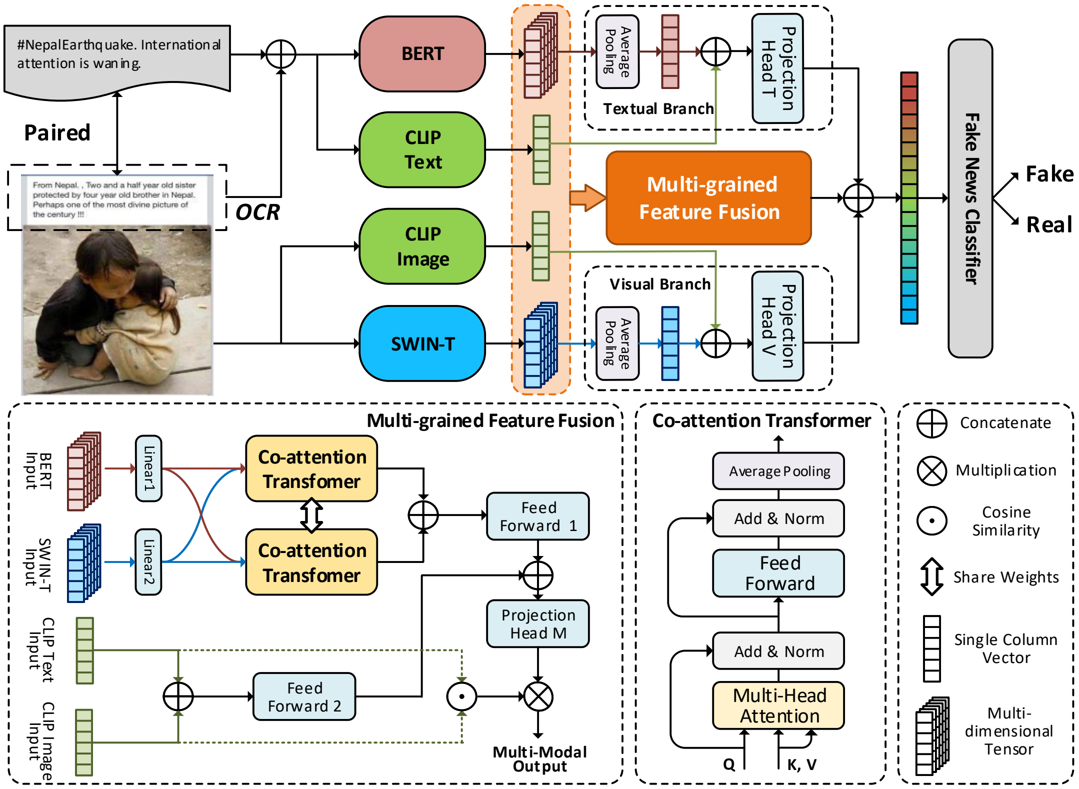


Figure 2: The architecture of the MMFN. BERT, Swin-T, and CLIP are used to encoded the features of different modalities of multi-modal news. Encoded features of different granularities are fused through CT and projection heads. CLIP similarity score is calculated to weight different modal features adaptively for the classifier to classify fake news.

making it suitable for fake news detection even in zero-shot scenarios [27]. Many works [13, 18] have shown CLIP's great ability in generalizing to unknown fields. Therefore, we use CLIP multi-modal features to enrich the global correlation information of textual and visual features. Given the multi-modal news $\mathbf{X} = \{\mathbf{T}, \mathbf{V}\}$, we denote the CLIP-encoded features as $\mathbf{X}^c = [t^c, v^c]$, where $t^c, v^c \in \mathbb{R}^{d_c}$ are two vectors of length d_c.

3.2 Multi-grained Feature Fusion

Fine-Grained Fusion via Transformer. Since BERT and Swin-T are not multi-modal models, there is a large gap between the features they extract, which cannot directly realize information interaction. To effectively fuse the textual and visual features of posts, we use a CT [20] to achieve information inter-modal complementarity. As shown in Figure 2, CT consists of a multi-headed attention network and a feed forward neural network, both followed by a residual connection and a layer normalization.

We denote different modal inputs as I_1 and I_2, respectively. In CT, I_1 is used as queries Q, and I_2 is used as keys K and values V. CT computes the co-attention matrix of each head as:

$$h_i = Softmax(\frac{(I_1 W_i^q)(I_2 W_i^k)^T}{d^h}) I_2 W_i^v, \tag{1}$$

where the projection matrices $W_i^q, W_i^k, W_i^v \in \mathbb{R}^{d_m \times d_h}$, $d_h = d_m/m$, and d_m is the dimension of the CT model, while m is the number of heads.

The multi-head attention is the concatenation of all co-attention matrices following a projection matrix:

$$H = (h_1; h_2; \cdots; h_m) W^o, \tag{2}$$

where symbol ; represents the concatenate operation and $W^o \in \mathbb{R}^{d_m \times d_m}$.

After H and I_1 pass through the FFN with two layer normalization, an attention-based multi-modal representation H' is obtained:

$$H' = \mathrm{Norm}(I_1 + \mathrm{FFN}(\mathrm{Norm}(I_1 + HI_1))). \tag{3}$$

Finally, the multi-modal representation H' is average pooled to a feature vector F as the output of CT:

$$F = \mathrm{CT}(I_1, I_2) = \overline{H'}. \tag{4}$$

In our model, after the inputs BERT and Swin-T features $\mathbf{T}^b, \mathbf{V}^s$ are mapped to the same dimension $\mathbb{R}^{d_m}$ through linear layers, they are input into a shared weighted CT in different front and rear order as I_1 and I_2 to obtain the output features, a visual attention weighted textual feature F^{vt} and a textual attention weighted visual feature F^{tv}, respectively:

$$\begin{cases} F^{vt} = \mathrm{CT}((\mathbf{T}^b W^t), (\mathbf{V}^s W^v)) \\ F^{tv} = \mathrm{CT}((\mathbf{V}^s W^v), (\mathbf{T}^b W^t)) \end{cases}, \tag{5}$$

where $W^t \in \mathbb{R}^{d_b \times d_m}$ and $W^v \in \mathbb{R}^{d_s \times d_m}$.

MCAN [38] and HMCAN [26] also use Transformer to fuse features. There are structural and functional differences between the CT in MMFN and the Transformers in HMCAN and MCAN. MCAN takes the pooled feature vectors as the input to the stacked co-attention blocks, which outputs a coarse-grained modality correlation vector; while the attention structure of MMFN is calculated at the token level and outputs fine-grained fused features. The Transformer in HMCAN consists of a single-head self-attention to enhance intra-modal information and another one to capture inter-modal information; considering BERT and Swin-T already use self-attention to achieve intra-modal interaction, to avoid impacting pre-training model and alleviate the overfitting problem, CT of MMFN only includes a multi-head co-attention to achieve inter-modal interaction. In addition, the Transformers in HMCAN do not share weights, while Transformers of MMFN share weights to make CT have the function of modal alignment.

Coarse-grained Fusion via CLIP and Multi-modal Representation Generation. We fuse the CLIP-encoded features as robust coarse-grained features that reflect the global semantic correlation information. It can be assumed that the outputs t^c and v^c of the CLIP encoders have effectively eliminated the inter-modality gap through contrastive learning. This allows subsequent network learning to effectively utilize the information from different modalities.

Specifically, t^c and v^c are concatenated and fed into a feed forward neural network with a linear layer, a batch norm layer, and a *ReLU* activation function. Additionally, F^{vt} and F^{tv} are also concatenated and fed into another feed forward neural network with the same architecture. Thus, after two feed forward networks, we have the fused multi-modal fine-grained feature M^f and the coarse-grained feature M^c:

$$\begin{cases} M^f = \mathrm{FFN}_1(F^{vt}; F^{tv}) \\ M^c = \mathrm{FFN}_2(t^c; v^c) \end{cases} . \tag{6}$$

Finally, the multi-modal features M^f and M^c are concatenated and fed into a projection head Φ_M to generate the multi-modal representation. Each element in the output vector will be multiplied by a similarity score, which will be introduced in the next paragraph. The generated multi-modal representation is denoted as F^m:

$$F^m = \mathit{similarity} \cdot \Phi_M(M^f; M^c). \tag{7}$$

3.3 Uni-modal Branches and Modality Weighting via CLIP Similarity

Multi-modal fused features generally reflect the correlation information between the two modalities, which is easily affected by ambiguity. To solve the problem that the feature representation ability of multi-modal fusion decreases when the modal is of high ambiguity, we designed a uni-modal textual branch and a uni-modal visual branch respectively.

For the textual branch, we pool the BERT features into a feature vector on the dimension of token level, concatenate it with the CLIP-text feature vector, then pass it through a projection head consisting of two fully connected networks with *ReLU* activation functions to get the uni-modal textual representation; similarly, for the visual branch, we concatenate the pooled Swin-T features with the CLIP-image feature, and obtain the uni-modal visual through a projection head mapping with the same structure but different parameters as the textual branch's:

$$\begin{cases} F^t = \Phi_T(\overline{\mathbf{T}^b}; t^c) \\ F^v = \Phi_V(\overline{\mathbf{V}^s}; v^c) \end{cases} , \tag{8}$$

where Φ_T and Φ_V are the projection heads.

If we directly send the uni-modal branch representations to the classifier for making the decision, the classifier may be more inclined to use the multi-modality representation with deeper network to fit the results, while the uni-modal branch could interfere with the decision and cause more serious ambiguity problems. To overcome such limitations, inspired by CAFE, we use the CLIP cosine similarity as a coefficient weighting the multi-modal feature to guide the classifier's learning process. The cosine similarity is calculated as follows:

$$\mathit{similarity} = \frac{t^c \cdot (v^c)^T}{\|t^c\| \, \|v^c\|}. \tag{9}$$

3.4 Fake News Classifier

After obtaining the fused multi-modal representation, uni-modal text representation, and uni-modal visual representation, we concatenate them as the input to a classifier and get the output $\hat{y}$ representing the probability of a news being fake:

$$\hat{y} = \mathrm{FNC}(F^m; F^t; F^v), \tag{10}$$

where $\mathrm{FNC}(\cdot)$ is the fake new classifier consisting of a two-layer fully connected network with *ReLu* activation function.

The objective function is to minimize the cross-entropy loss to correctly predict the real and fake news.

$$\mathcal{L} = y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}), \tag{11}$$

where y is the ground truth label.

4 EXPERIMENTS

4.1 Experimental Setup

Dataset. We use three public real-world datasets collected from social media, namely, Weibo [14], Twitter [6], and Gossipcop [29].

Weibo is a widely used Chinese dataset in fake news detection. The real news was collected from Xinhua News Agency, an authoritative news source of China. In experiments, the uni-modal news posts with no image or no text description were filtered out. The training set contains 3, 783 real news and 3, 675 fake news, and the test set contains 1,685 news.

The Twitter dataset was released for MediaEval Verifying Multimedia Use task [6] and is also a well-known multi-modal dataset for fake news detection. In experiments, following existing works we filter the tweets with videos attached and the non-English tweets. After filtering, the training set contains 4, 031 real news and 5, 139 fake news, and the test set contains 1, 406 posts.

Gossipcop dataset is a English full-length article news dataset collected from the entertainment domain of FakeNewsNet [29] repository. Gossipcop contains 10, 010 training news, including 7, 974 real news and 2, 036 fake news. The test set has 2, 285 real news and 545 fake news.

Implementation Details. In our experiments, we set $d_m = 512$ and $m = 8$ for the CT. The textual embedding dimension of BERT

Table 1: Performance comparison between MMFN and state-of-the-art methods on Weibo, Twitter, and Gossipcop datasets. The best performance is highlighted in bold.

	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F1-score	Precision	Recall	F1-score
Weibo	Spotfake [32]	0.892	0.902	**0.964**	**0.932**	0.847	0.656	0.739
	CAFE [7]	0.840	0.855	0.830	0.842	0.825	0.851	0.837
	MCAN [38]	0.899	0.913	0.889	0.901	0.884	0.909	0.897
	LIIMR [31]	0.900	0.882	0.823	0.847	0.908	**0.941**	**0.925**
	HMCAN [26]	0.885	0.920	0.845	0.881	0.856	0.926	0.890
	MMFN	**0.923**	**0.921**	0.926	0.924	**0.924**	0.920	0.922
Twitter	Spotfake [32]	0.777	0.751	**0.900**	0.820	0.832	0.606	0.701
	CAFE [7]	0.806	0.807	0.799	0.803	0.805	0.813	0.809
	MCAN [38]	0.809	0.889	0.765	0.822	0.732	0.871	0.795
	LIIMR [31]	0.831	0.836	0.832	0.830	0.825	0.830	0.827
	HMCAN [26]	0.897	**0.971**	0.801	0.878	0.853	0.979	0.912
	MMFN	**0.935**	0.960	0.856	**0.905**	**0.924**	**0.980**	**0.951**
Gossipcop	SAFE [44]	0.838	0.758	0.558	0.643	0.857	0.937	0.895
	Spotfake+ [30]	0.858	0.732	0.372	0.494	0.866	0.962	0.914
	DistilBert [2]	0.857	**0.805**	0.527	0.637	0.866	0.960	0.911
	CAFE [7]	0.867	0.732	0.490	0.587	0.887	0.957	0.921
	MMFN	**0.894**	0.799	**0.598**	**0.684**	**0.910**	**0.964**	**0.936**

is set to $d_b = 768$, and we use the “bert-base-chinese” model for Chinese data and the “bert-base-uncased” model for English data. The input text length is set to 300 words, i.e., $n_w = 300$. For the Swin-T, we use the “swin-base-patch4-window7-224” model to encode visual features and set the input image size to 224 × 224. The number of patches in Swin-T is $n_p = 49$ and the dimension of the visual embedding is $d_s = 1024$. The input image size for CLIP is also set to to 224 × 224. Science CLIP has not pre-trained Chinese text model, we use Google Translation API [15] to translate Chinese texts to English. We use a summary generation model [28] to generate summary statements for texts longer than 50 words to meet the input size requirements of CLIP. The used pre-trained CLIP model is “ViT-B/32” with feature dimension of $d_c = 512$. We fine-tune BERT and Swin-T during the training stage, while freezing the parameters of CLIP due to its difficulty in training on small datasets. Similar to MCAN [38], we freeze the BERT model on the Twitter dataset to alleviate overfitting. The feed forward network 1 and the feed forward network 2 have hidden size of 256, and the projection heads had hidden units of 256 and 16, respectively. The hidden sizes of the two fully conneced layers in the classifier are 48 and 2, respectively. The batch size is set to 16. The dataset was pre-processed to remove all invalid messages after "@", "#", and "http:". The Adam optimizer [17] was utilized with the default parameters. Aside from fine-tuning BERT and Swin Transformer using a learning rate of 1×10^{-5}, the learning rate for the network was set to 1×10^{-3}. The model is trained for 100 epochs with early stopping to prevent over-fitting.

4.2 Performance Comparison

We compare the performance of MMFN with other state-of-the-art methods and the comparison results are presented in Table 1. The evaluation metrics are accuracy, precision, recall, and F1-score, which are commonly used to measure the performance of a binary classification problem. As shown in Table 1, MMFN outperforms the other methods across all three datasets in terms of Accuracy. Specifically, MMFN achieves the highest accuracy of **92.3%**, **93.5%**, and **89.4%**, respectively, on the three real-world datasets, surpassing the state-of-the-art method by **2.3%**, **3.8%**, and **2.7%**. In terms of precision, recall, and F1 score, MMFN ranks either first or second on nearly all tests, showcasing its effectiveness.

The compared methods, SAFE, Spotfake, DistilBert and Spotfake+ have limitations in their methods for fake news detection. SAFE learns similarity between text and visuals, but may misclassify real posts with weak correlation as fake news due to ignoring ambiguity. DistilBert guides the detection by checking the effect of user-related constraints on article latent space, yet it neglects visual information of news thus lead to less competitive performance. Spotfake and Spotfake+ simply concatenate textual and visual representations without adequate cross-modal interaction and fusion, leading to suboptimal performance. CAFE defines and utilizes cross-modal ambiguity to alleviate the problem of disagreement between different modalities. It achieves better experimental results than Spotfake, Spotfake+, and SAFE on the Twitter and Gossipcop datasets.

LIIMR achieves improved results on the Weibo dataset due to its ability to capture fine-grained salient image and text features. The decent experimental results of MCAN and HMCAN prove the effectiveness of the Transformer-based multi-modal fusion network. However, these approaches concentrate only on fine-grained feature mining, neglecting the coarse-grained information that provides insight into global semantics.

The superiority of MMFN over the other methods can be attributed to three factors. 1) The Swin-T component is capable of extracting fine-grained features that complement the features generated by the BERT encoder. Additionally, the pre-trained CLIP encoder is capable of generating coarse-grained text and image features that possess rich semantic information within a shared

Table 2: Ablation study on modalities, granularities, and architecture designs of MMFN on Weibo, Twitter, and Gossipcop datasets. The best performance is highlighted in bold.

	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F1-score	Precision	Recall	F1-score
Weibo	MMFN w/o T	0.825	0.846	0.815	0.830	0.803	0.836	0.819
	MMFN w/o V	0.901	0.859	0.940	0.898	0.944	0.868	0.904
	MMFN w/o F	0.818	0.793	0.839	0.815	0.844	0.800	0.821
	MMFN w/o C	0.909	0.914	0.907	0.911	0.904	0.912	0.908
	MMFN w/o CT	0.905	0.849	**0.959**	0.900	**0.963**	0.861	0.909
	MMFN w/o U	0.912	**0.939**	0.891	0.915	0.884	**0.934**	0.908
	MMFN w/o W	0.906	0.921	0.895	0.908	0.890	0.917	0.903
	MMFN	**0.923**	0.921	0.926	**0.924**	0.924	0.920	**0.922**
Twitter	MMFN w/o T	0.933	0.963	0.850	0.903	0.919	0.981	0.949
	MMFN w/o V	0.739	0.691	0.580	0.630	0.762	0.839	0.799
	MMFN w/o F	0.866	**0.993**	0.708	0.826	0.805	**0.996**	0.891
	MMFN w/o C	0.882	0.817	0.817	0.817	0.913	0.913	0.913
	MMFN w/o CT	0.900	0.965	**0.885**	**0.923**	0.793	0.932	0.857
	MMFN w/o U	0.918	0.929	0.835	0.880	0.913	0.965	0.938
	MMFN w/o W	0.915	0.960	0.810	0.879	0.893	0.979	0.934
	MMFN	**0.935**	0.960	0.856	0.905	**0.924**	0.980	**0.951**
Gossipcop	MMFN w/o T	0.836	0.702	0.255	0.374	0.846	0.974	0.905
	MMFN w/o V	0.888	0.818	0.536	0.648	0.898	0.972	0.933
	MMFN w/o F	0.862	0.691	0.517	0.592	0.891	0.945	0.917
	MMFN w/o C	0.885	**0.844**	0.495	0.624	0.890	**0.978**	0.932
	MMFN w/o CT	0.888	0.756	**0.615**	0.678	**0.912**	0.953	0.932
	MMFN w/o U	0.885	0.799	0.539	0.644	0.898	0.965	0.932
	MMFN w/o W	0.889	0.764	0.611	0.679	0.911	0.968	0.933
	MMFN	**0.894**	0.799	0.598	**0.684**	0.910	0.964	**0.936**

semantic space. This allows for complementary feature representation at both fine and coarse granularities. 2) The CT component enables inter-modal interaction at the token level, thus facilitating the fine-grained fusion of multiple modalities. 3) The utilization of uni-modal branches with CLIP-based weighting effectively mitigates the issue of ambiguity.

4.3 Ablation Studies

We evaluate the impact of the key components in MMFN on its performance by conducting experiments with various and partial configurations of the model. For each experiment, we remove a different component and retrain the model from scratch. The compared variants of MMFN are implemented as follows:

1) MMFN w/o T. The text-related modules are removed and only the uni-modal visual features encoded by Swin-T and CLIP image coder are used.

2) MMFN w/o V. The visual-related modules are removed, and only the uni-modal textual features encoded by BERT and CLIP text coder are retained.

3) MMFN w/o F. The BERT-related and Swin-T-related modules are removed, and the fine-grained features are not utilized. Instead, the CLIP-coded text and image features are directly concatenated into a multi-modal representation, and the two CLIP features are used as two separate uni-modal representations.

4) MMFN w/o C. The CLIP-related modules are removed, and the coarse-grained features are not employed. The classification process is performed using only fine-grained features.

5) MMFN w/o CT. The CT module is removed, and the features encoded from BERT and Swin-T are directly concatenated.

6) MMFN w/o U. The textual and visual uni-modal branches are removed, and only the multi-modal fused representation is used for classification.

7) MMFN w/o W. The CLIP-weighting modules are removed, and the multi-modal fused feature is not weighted.

Contributions from Different Modalities. To assess the contributions of different modalities to the overall performance of MMFN model, we compare the results of MMFN w/o T and MMFN w/o V with the completed MMFN model. The results indicate that the performance of MMFN decreases when either the visual or textual modality is absent. This suggests that both modalities are crucial for the final performance of the model.

Additionally, we found that the performance of MMFN w/o T is poor on the Weibo and Gossipcop datasets, but second-best on the Twitter dataset. On the other hand, MMFN w/o V performs poorly on the Twitter dataset, but outperforms the MMFN w/o T and some multi-modal ablated models on Weibo and Gossipcop. These observations indicate that: 1) The distribution of news on different datasets and different social media platforms is diverse. For instance, the text of Weibo news contains rich clues for fake news detection, while Twitter users tend to express information through images. Meanwhile, the main information in long news on Gossipcop is located in the text, while images play a minor role. This implies that uni-modal models may have difficulties adapting to different social environments. 2) The uni-modal models outperform several of the ablated multi-modal models, suggesting that multi-modal learning is effective only when the cue information from different modalities is adequately mined and integrated using an effective mechanism.

Influence of Different Granularities. To analyze the effect of different granularities on the performance of the model, we compare the results of MMFN w/o C and MMFN w/o F with the completed MMFN model. The results show that both coarse-grained and fine-grained features contribute to the performance of the model. The performance of the model decreases more when the fine-grained features are removed than when the coarse-grained features are removed. This suggests that fine-grained features play a more important role in the final performance of the model. The suggestion follows our intuition, and mining more fine-grained features to detect fake news is the direction of the recent works like LIIMR and HMCAN, etc. However, the results of the ablation study shows that the coarse-grained features is useful to assist the fine-grained features to improve the detection ability.

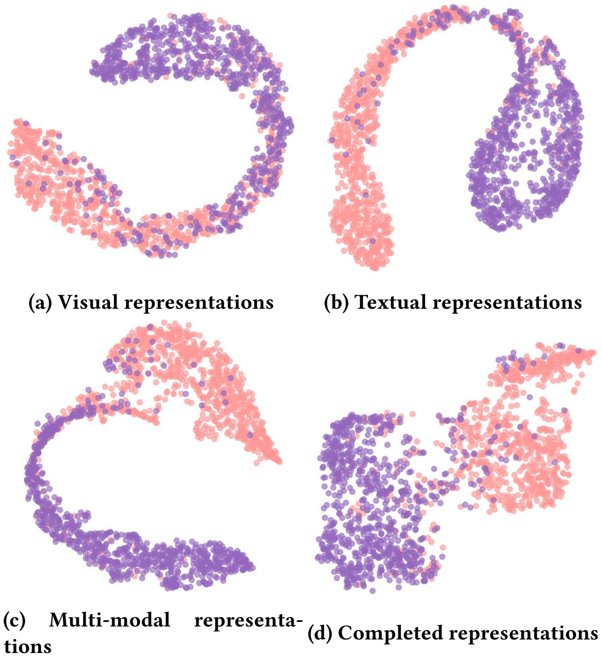
Effectiveness of Each Component. To evaluate the effectiveness of each component in the MMFN model, we compare the results of the completed MMFN model with the ablated models MMFN w/o CT, MMFN w/o U, MMFN w/o W. As can be seen in Table 2, the following observations are made: 1)The integration of visual and textual modalities through the CT module enhances the representation capability and fine-grained feature fusion, which is reflected in the improved performance of the complete model compared to the MMFN w/o CT. 2) The Completed MMFN model outperforms MMFN w/o U, demonstrating the effectiveness of the uni-modal branch designed to handle ambiguity in improving the performance. 3) The results also show that both the complete MMFN model and MMFN w/o U outperform MMFN w/o W on the Weibo and Twitter datasets, indicating that weighting the fusion features helps alleviate ambiguity. On the other hand, simply adding uni-modal branches without weighting may harm performance. Overall, the results suggest that the full combination of components is necessary for achieving the best performance of the MMFN model.

4.4 T-SNE Visualizations

In Figure 3, we further analyze the separability of the different modal representations learned by MMFN and its variants on Weibo using t-SNE [33] visualizations. The visual, textual, multi-modal, and completed representations are obtained from MMFN w/o T, MMFN w/o I, MMFN w/o U, and completed MMFN, respectively. The dots of the same color indicate instances of the same label.

The visual representations learned by MMFN w/o T have some overlap with those of different labels, suggesting that the visual information alone is not sufficient for classification on Weibo. On the other hand, the textual representations learned by MMFN w/o I show a relatively better separability compared to the visual representations, indicating that the textual information can provide some discriminative information for the classification task.

The multi-modal representations learned by MMFN w/o U are more separable compared to the visual representations and are slightly better in separability to the textual representations. This indicates that the combination of textual and visual information can provide complementary information and improve the separability of the representations. However, due to the absence of uni-modal branches, MMFN w/o U isn't as effective as the completed model. Finally, the representations learned by the completed MMFN show the best separability among all the variants, surpassing the separability of multi-modal representations. This suggests that the proposed method is effective in handling the usage of multi-modal representations to addressing the ambiguity problem, which significantly enhances the separability.


(a) Visual representations (b) Textual representations

(c) Multi-modal representations (d) Completed representations

Figure 3: T-SNE visualizations of the different modal representations on the test dataset of Weibo.

Overall, these results demonstrate not only the importance of combining both visual and textual information for fake news detection on Weibo, but also the effectiveness of the proposed fusion method in organizing multi-grained multi-modal representations.

5 CONCLUSIONS

In this paper, we present a novel multi-modal fake news detection method called MMFN, which uses BERT and Swin-T with a Transformer to obtain fine-grained multi-modal feature and uses CLIP to acquire coarse-grained multi-modal feature. In addition, we introduce uni-modal branches with CLIP similarity weighting to adaptively assist multi-modal classification. We conduct comprehensive experiments on two well-known datasets for multi-modal fake news detection. The results show that MMFN outperforms many of the state-of-the-art methods.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China under Grants U20B2051, 62072114, U22B2047 and U1936214.

REFERENCES

[1] Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. 2022. Open-Domain, Content-based, Multi-modal Fact-checking of Out-of-Context Images via Online Resources. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14940–14949.

[2] Liesbeth Allein, Marie-Francine Moens, and Domenico Perrotta. 2021. Like Article, Like Audience: Enforcing Multimodal Correlations for Disinformation Detection. *arXiv preprint arXiv:2108.13892* (2021).

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2018), 423–443.

[4] Gaurav Bhatt, Aman Sharma, Shivam Sharma, Ankush Nagpal, Balasubramanian Raman, and Ankush Mittal. 2018. Combining neural, statistical and external features for fake news stance identification. In *Proceedings of the The Web Conference*. 1353–1357.

[5] Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 549–556.

[6] Christina Boididou, Symeon Papadopoulos, Markos Zampoglou, Lazaros Apostolidis, Olga Papadopoulou, and Yiannis Kompatsiaris. 2018. Detection and visualization of misleading content on Twitter. *International Journal of Multimedia Information Retrieval* 7, 1 (2018), 71–86.

[7] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal Ambiguity Learning for Multimodal Fake News Detection. In *Proceedings of The Web Conference*. 2897–2905.

[8] Marcos V Conde and Kerem Turgutlu. 2021. CLIP-Art: contrastive pre-training for fine-grained art classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3956–3960.

[9] Nadia K Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information science and Technology* 52, 1 (2015), 1–4.

[10] Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. 2021. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1574–1583.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[13] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921* (2021).

[14] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the ACM international conference on Multimedia*. 795–816.

[15] Gregory Johnson. 2012. Google Translate http://translate. google. com. *Technical Services Quarterly* 29, 2 (2012), 165–165.

[16] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *Proceedings of the The Web Conference*. 2915–2921.

[17] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[18] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. 2022. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546* (2022).

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10012–10022.

[20] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems* 32 (2019).

[21] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 3818–3824.

[22] Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. MDFEND: Multi-domain Fake News Detection. In *Proceedings of the ACM International Conference on Information & Knowledge Management*. 3343–3347.

[23] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).

[24] Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. 2021. Improving Fake News Detection by Using an Entity-enhanced Framework to Fuse Diverse Multimodal Clues. In *Proceedings of the ACM International Conference on Multimedia*. 1212–1220.

[25] Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. 2019. Exploiting multi-domain visual information for fake news detection. In *Proceedings of the IEEE International Conference on Data Mining*. IEEE, 518–527.

[26] Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 153–162.

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. 8748–8763.

[28] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* (2019).

[29] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data* 8, 3 (2020), 171–188.

[30] Shivangi Singhal, Anubha Kabra, Mohit Sharma, Rajiv Ratn Shah, Tanmoy Chakraborty, and Ponnurangam Kumaraguru. 2020. Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13915–13916.

[31] Shivangi Singhal, Tanisha Pandey, Saksham Mrig, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2022. Leveraging Intra and Inter Modality Relationship for Multimodal Fake News Detection. In *Companion Proceedings of The Web Conference 2022*. 726–734.

[32] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *Proceedings of the IEEE international conference on multimedia big data*. 39–47.

[33] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 11 (2008).

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[35] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multimodal fake news detection. In *Proceedings of the ACM international conference on knowledge discovery & data mining*. 849–857.

[36] Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2020. Fake news detection via knowledge-driven multimodal graph convolutional networks. In *Proceedings of the International Conference on Multimedia Retrieval*. 540–547.

[37] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Zhentao Tan, Lu Yuan, Weiming Zhang, and Nenghai Yu. 2021. Hairclip: Design your hair by text and reference image. *arXiv preprint arXiv:2112.05142* (2021).

[38] Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal Fusion with Co-Attention Networks for Fake News Detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2560–2569.

[39] Junxiao Xue, Yabo Wang, Yichen Tian, Yafei Li, Lei Shi, and Lin Wei. 2021. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management* 58, 5 (2021), 102610.

[40] Keren Ye and Adriana Kovashka. 2021. A case study of the shortcut effects in visual commonsense reasoning. In *Proceedings of the AAAI conference on Artificial Intelligence*, Vol. 35. 3181–3189.

[41] Qichao Ying, Xiaoxiao Hu, Yangming Zhou, et al. 2023. Bootstrapping Multi-view Representations for Fake News Detection. In *AAAI*.

[42] Feng Yu, Qiang Liu, Shu Wu, Liang Wang, Tieniu Tan, et al. 2017. A Convolutional Approach for Misinformation Identification.. In *IJCAI*. 3901–3907.

[43] Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of The Web Conference 2021*. 3465–3476.

[44] Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: Similarity-Aware Multimodal Fake News Detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 354–367.

[45] Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. 2022. Multimodal fake news detection via CLIP-guided learning. *arXiv preprint arXiv:2205.14304* (2022).

[46] Yongchun Zhu, Qiang Sheng, Juan Cao, Qiong Nan, Kai Shu, Minghui Wu, Jindong Wang, and Fuzhen Zhuang. 2022. Memory-Guided Multi-View Multi-Domain Fake News Detection. *IEEE Transactions on Knowledge and Data Engineering*

(2022).

[47] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)* 51, 2 (2018), 1–36.