Information Processing and Management 60 (2023) 103120

Contents lists available at ScienceDirect

Information Processing and Management

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

Multimodal fake news detection via progressive fusion networks

Jing Jing, Hongchen Wu *, Jie Sun, Xiaochang Fang, Huaxiang Zhang

School of Information Science and Engineering, Shandong Normal University, Shandong, 250014, China

ARTICLE INFO

Keywords:
Fake news detection
Multimodal fusion
Social media
Neural network

ABSTRACT

Multimodal fake news detection methods based on semantic information have achieved great success. However, these methods only exploit the deep features of multimodal information, which leads to a large loss of valid information at the shallow level. To address this problem, we propose a progressive fusion network (MPFN) for multimodal disinformation detection, which captures the representational information of each modality at different levels and achieves fusion between modalities at the same level and at different levels by means of a mixer to establish a strong connection between the modalities. Specifically, we use a transformer structure, which is effective in computer vision tasks, as a visual feature extractor to gradually sample features at different levels and combine features obtained from a text feature extractor and image frequency domain information at different levels for fine-grained modeling. In addition, we design a feature fusion approach to better establish connections between modalities, which can further improve the performance and thus surpass other network structures in the literature. We conducted extensive experiments on two real datasets, Weibo and Twitter, where our method achieved 83.3% accuracy on the Twitter dataset, which has increased by at least 4.3% compared to other state-of-the-art methods. This demonstrates the effectiveness of MPFN for identifying fake news, and the method reaches a relatively advanced level by combining different levels of information from each modality and a powerful modality fusion method.

1. Introduction

During the time of global online community recovery and combat against infodemic, social media platforms such as Twitter, Weibo, and other social applications are still major channels for people to obtain massive information; meanwhile, people may publish and forward fake news on social media websites at very low cost. Fake news fully caters to the audience's curiosity, which leads to its rapid spread. In recent decades, major unexpected events have occurred frequently, thereby leading to the proliferation and spread of fake news, infodemic seriously disturbed the social order and caused social panic. For example, in the 2016 U.S. presidential election (Bovet & Makse, 2019), fake news on social media was proliferating, which severely affected citizens' voting intentions on political choices and, thus, the fairness of the election process and results will also be affected. During the COVID-19 epidemic (Satu, et al., 2021), the Internet was flooded with fabricated misinformation, such as that salt water, tea, and vinegar could be effective treatment against COVID-19 pandemic, causing great difficulties in figuring ground truth information in online community.

Currently, social media is facing substantial threats and challenges, and articles with images and text are becoming increasingly popular on most searched hashtags in social media. This study considers fake news that contains media content of multiple modalities, e.g., images and text. Compared with text-only articles, images contain richer information and can better express and

* Corresponding author.
E-mail address: wuhongchen@sdnu.edu.cn (H. Wu).

https://doi.org/10.1016/j.ipm.2022.103120
Received 7 May 2022; Received in revised form 12 September 2022; Accepted 12 October 2022
Available online 29 October 2022

(a) (b) (c)

Fig. 1. Fake news sample published by Weibo and Twitter: (a) The urban village near Huanggang Port undergoing renovation and demolition. (b) Every time a hurricane sweeps off the US coast, sharks swim up the street. (c) Car thieves were electrocuted by electric vehicles, and the family of a deceased thief demanded 200,000 RMB in compensation.

spread content, thereby attracting the attention of more readers. Research study (Jin, Cao, Zhang, Zhou, & Tian, 2016) have found that the average number of retweets for articles with images is 11 times higher than for articles without images. However, the ease of access to news has also enabled such fake news to spread widely. Fig. 1 shows three examples of fake news from the Weibo dataset, where each example contains text and additional images. There is little evidence in the text in the left section that this is fake news, but the images are clearly fake. The middle example indicate a fake news candidate from the perspective of both the text and the image. In the example on the right, the image looks normal, while the text content suggests that it may be fake news. One conclusion that is drawn from these examples is that combining text and additional images is more beneficial for detecting fake news. Therefore, visual content shall not be ignored because it is an essential aspect in detecting fake news, and thus, it is necessary to establish an automatic method for detecting the authenticity of articles with images to mitigate the severely negative impact of fake news.

Many researchers focus on finding solutions for fake news detection, such as manual fact-checking, for which the methods include Zhou, Zafarani, Shu, and Liu (2019) expert fact-checking and crowdsourced fact-checking methods. Expert fact-checking is highly accurate but time-consuming and labor-intensive. Crowdsourced fact-checking is highly scalable, but the verification accuracy is not high. Due to the limitations of manual fact-checking methods, researchers have used expert knowledge to manually extract features from news text content and trained fake news classifiers using traditional machine learning algorithms (Kwon, Cha, Jung, Chen, & Wang, 2013; Yang, Liu, Yu, & Yang, 2012), but this approach lacks comprehensiveness and flexibility. The established deep learning models have achieved relatively good performance due to their strong feature extraction capability, which enables them to automatically extract news features from news content. Chen, Zhou, Trajcevski, and Bonsangue (2022) proposed a method applying user multi-view learning and attention for rumor detection. This method can better learn the representations of different views of users in the propagation path of tweets and can fuse the learned representations through a fusion mechanism. Chi and Liao (2022) proposed QA-AXDS based on quantitative argumentation to detect rumors and interaction with the user through an explanation model in the form of a dialogue tree, thus helping the user to provide explanations about the results.

As fake news becomes more diverse, the authenticity of articles with images poses higher requirements and greater challenges for fake information detection techniques, and various deep learning-based methods have been successfully applied to multimodal fake news detection. First, some models, such as that of Khattar, Goud, Gupta, and Varma (2019), used multimodal variational encoders to simply extract and fused features from text and images, but the components that they used to capture multimodal contexts were too simple to extract higher-order complementary information from multimodal contexts. Second, an end-to-end network was established by Jin, et al. (2017) with a fake news detection model that was designed using an RNN that utilized a local attention mechanism in combination with textual images and social context features, and Wang, et al. (2018) built an event adversarial neural network (EANN) that used an event discriminator to learn feature representations of text and images in articles, but adding additional auxiliary features would increase the cost of detection. A multimodal fake news detection framework named CARMN was proposed by Song, Ning, Zhang, and Wu (2021b) from the perspective of multichannel convolution neural networks. However, these methods only considered the spatial domain of the image and neglected the frequency domain information of the image (Qi, Cao, Yang, Guo, & Li, 2019). Third, (Wu, Zhan, Zhang, Wang, & Xu, 2021) proposed multimodal co-attention networks (MCAN) for fake news detection. MCAN can learn the interdependence between multimodal features and achieve better results in fake news detection. But MCAN focused only on the role of deep-level features of each modality and ignored the utilization of shallow-level features.

In summary, our research objectives focus on exploring a fusion strategy that make full use of modality features in different hierarchical spaces, that improve the perception of the features of different hierarchical spaces, and that alleviate the suppression of important information. This study fills the gap between the previous works, whose generic approaches utilized fake news detection to process information only from different modalities through encoding and pooling operations, making the suppression of important information and learning progress of the fine-grained complementary information among modalities rather difficult. In addition, our designed fusion module enables finer-grained fusion of information from each modality.

Specifically, we propose a fake news detection model of progressive multimodal fusion for determining the authenticity of multimodal news, namely, MPFN, which models different modal information separately, extracts different levels of information, progressively fuses these levels of information from shallow to deep, and fuses different levels of multimodal information using a complex fully connected fusion mechanism that we designed. Our proposed model consists of three fundamental components: a text feature extractor, a visual feature extractor, and a progressive multimodal feature fusion process. First, text features are extracted by

coding text features using a pretrained BERT model (Devlin, Chang, Lee, & Toutanova, 2018). Second, the visual feature extractor, which consists of a Swin Transformer (Liu, et al., 2021) and VGG19 (Simonyan & Zisserman, 2014), is used to extract features from the complete information of the image spatial domain and frequency domain, respectively. To solve the problem of the loss of shallow features in previous work, we designed a multilevel visual and text fusion strategy and designed the Mlp Mixer for fusing the obtained visual and text feature representations with fine granularity to establish close relationships between and within modes.

Our main contributions are summarized as follows:

- We propose a novel end-to-end learning technique that uses a progressive fusion strategy to capture the information of various levels of representations of each modality.
- We design a fusion module, namely, Mlp Mixer, that can effectively establish the dependence relationships between modes and fuse the features of each mode with fine granularity.
- We evaluate MPFN on both Weibo and Twitter datasets for determining the authenticity of misinformation. The results show the effectiveness of the fusion strategy that we employ by demonstrating quantitatively that our approach outperforms state-of-the-art models.

The remainder of this paper is organized as follows: In Section 2, we summarize previous work on fake news detection. In Section 3, we describe our proposed fake news detection method in detail. In Section 4, we describe the datasets, baseline, and experimental results of our experiments. Finally, we present the conclusions of this study and discuss potential research directions in Section 5.

2. Related work

In Bondielli and Marcelloni (2019), Kumar and Geethakumari (2014) fake information is defined as information in which the maker intentionally misleads the reader and can confirm that the result is fake through another source. Verifying the authenticity of the posted news on social media is still a remaining challenge in fake news detection research field. Due to the diversity of fake news and its evolving forms, the detection methods need updates from simple single-modal detection approaches to multimodal detection approaches accordingly. The rest part of this section introduce the development of fake news detection techniques in detailed manner. Furthermore, modeling multimodal news content is introduced to explore the feasibility of fake news detection methods.

2.1. Single-modal fake news detection

Textual features. A news article includes many types of content, such as text, images, and social scenarios. Initially, many methods were used to detect the authenticity of articles based on their text content, and the primary methods that were applied are on the basis of text feature extraction in statistical or semantic level, such as the number of paragraphs of the text (Volkova, Shaffer, Jang, & Hodas, 2017), lexical percentage (Bond, et al., 2017; Potthast, Kiesel, Reinartz, Bevendorff, & Stein, 2017), number of symbols (Castillo, Mendoza, & Poblete, 2011), writing style (Chen, Conroy, & Rubin, 2015) and language style (Feng, Banerjee, & Choi, 2012), which have been studied in the fake news detection literature. Then, traditional machine learning algorithms were used to detect fake news. Unfortunately, extracting features from text manually is not only difficult and time-consuming in design but also fails to take full advantage of the content of the text. To solve this problem, many researchers used deep learning methods to identify fake news because deep learning techniques have powerful representation learning capabilities. An article learning model with the representatives of text features in time series was proposed by Ma, et al. (2016), basing on a recurrent neural network. A convolutional neural network was applied in the model by Yu, et al. (2017) in examining deeper interactions between essential features, which was the first approach in extracting low-level features. All in all, recent literature on deep learning approaches outperform the machine learning results in experimental results.

Visual features. With the rise of multimodal forms of news content, research on using visual features for disinformation detection is rapidly increasing. Many studies (Jin, et al., 2016; Wu, Yang, & Zhu, 2015) have carefully investigated the accompanying images and image types, among other information, using visual features. However, the learning mode of these methods is simple and has substantial limitations. Qi, et al. (2019) proposed a convolutional neural network (CNN)-based model for capturing image patterns, which initially used CNN to extract frequency-domain patterns, subsequently used a recurrent neural network for semantic detection of photo authenticity, and finally used an attention mechanism for the fusion of image patterns and frequency domain patterns. The addition of visual features improves the performance of fake news detection, but there are limitations on the scenarios in which these models can be used.

2.2. Multimodal fake news detection

Fusing text and visual features is a potential perspective of enhancing fake news detection performance due to fully learning the semantic features of data. With the diversified development of disinformation articles, many researchers have studied the problem of multimodal fake detection. Giachanou, Zhang, and Rosso (2020) combined textual, visual and semantic information to exploit the complementarity between modalities for multimodal fake news detection. Kumari and Ekbal (2021) proposed a multimodal fake news detection framework that maximize the correlation between text and image features for better multimodal information representation. Wang, Ma, Wang, Jha, and Gao (2021) delivered a fake news detection framework called MetaFEND that

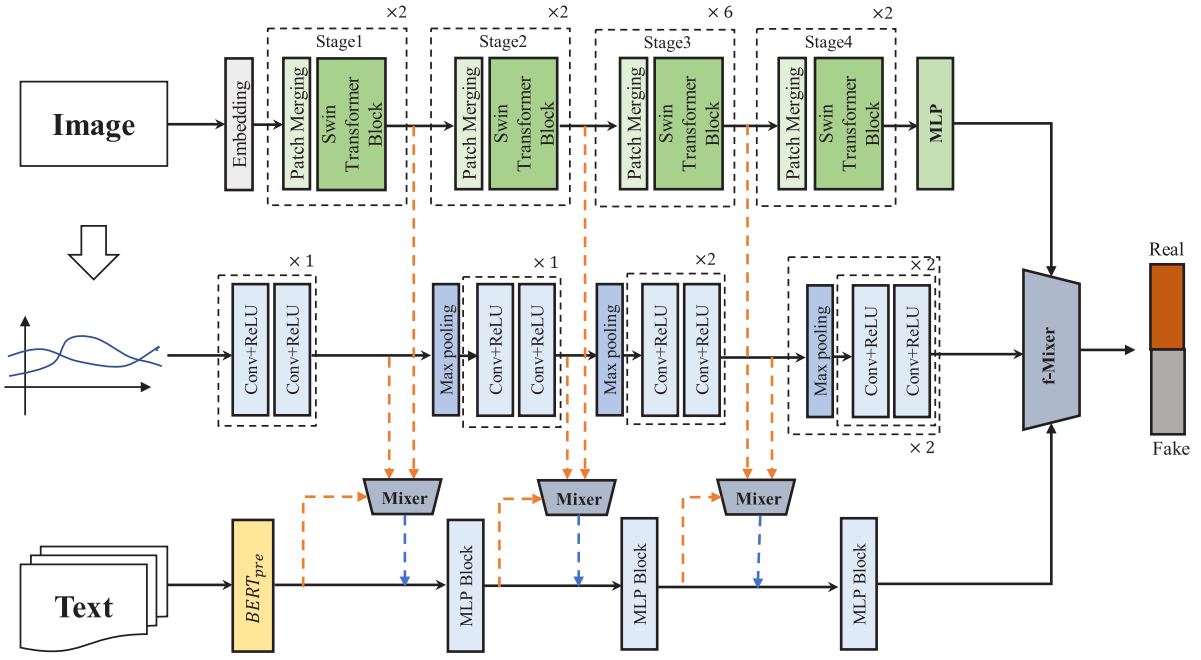

Fig. 2. Architecture of our MPFN model.

efficiently learns and performs breaking news detection based on a small number of verified posts. Tuan and Minh (2021) showed a method that can efficiently learn and fuse multimodal features in posts to detect fake news. Song, Ning, Zhang, and Wu (2021a) proposed a fake news detection framework KMGCN using feature representations made by fusing textual information, knowledge concepts and visual information with Graph Convolutional Network. Khattar et al. (2019) extracted modal information from text and images and performed simple fusion using multimodal variant encoders. A fake news detection model was designed with an RNN by Jin, et al. (2017), who created an end-to-end network combining text images and social context features for local attention mechanism. Giachanou et al. (2020) combined semantic representations obtained from image and article title similarity calculations with visual features of images learned by VGG-16 and text features learned by Transformers (BERT) for more effective fake news detection. Wang, et al. (2018) built an event adversarial neural network (EANN) that used event discriminators to learn feature representations of text and images in articles. As the complexity of fake news increases, various news media combine misleading or doctored images with text to mislead readers and enable rapid dissemination of articles, which requires us to consider the frequency domain information of images. However, many studies considered only the spatial domain information of pictures and ignored the changes in pictures in the frequency domain. With the significant improvement of generated image modeling based on CNN (Choi, Uh, Yoo, & Ha, 2020; Tran, Tran, Nguyen, Yang, & Cheung, 2019), it was difficult to distinguish fake images that were synthesized by a generated model from real images, which brought larger challenges to the fake information detection task. Zhang, Giachanou, and Rosso (2022) proposed SceneFND, a system combining text, contextual scenes, and visual representations, which used contextual scenes to integrate features and effectively improved the performance of fake news detection. Wu, et al. (2021) proposed multimodal co-attention networks (MCAN) for fake information detection, which combined the frequency domain information of images to learn the interdependencies between multimodal features. Error-level analysis (Ela) algorithm was performed by Xue, et al. (2021) for directly extracting the frequency domain from the images, in purpose of detecting the recompression features of forged images and the malicious stitching. Although the previous methods achieved better performance on the fake news detection task, most of them only considered the correlation of each modal information under deep features in the process of fusion of each modal information, while ignored the utilization of effective information of shallow features, limiting possible improvement of multimodal fake news detection performance.

To solve the above problems, we propose a new network model (MPFN) with a fusion strategy that makes full use of the spatial characteristics of different levels of multimodality to process and fully integrate information from different levels in multiple stages. Specifically, our designed module fuses the features between the modalities for enhancing their correlation and reached better performance.

3. Methodology

In this paper, we propose a progressive multimodal fusion fake news detection model for determining the authenticity of multimodal news. Fig. 2 illustrates our proposed progressive multimodal rumor detection model, which consists mainly of a BERT-based text feature extractor, a transformer-based visual feature extractor, and a progressive multimodal feature fusion process.

The fake news detection task focuses on checking the truthfulness of news on social media and can be seen as a binary classification problem. Given news with text and images, we first extract features from the spatial domain, frequency domain, and text using three different modules. Then, to enhance the fine-grained fusion between modalities, we learn multimodal information via a progressive fusion approach and, finally, fuse feature information between modalities via a more fine-grained approach by using the Mlp Mixer. The model labels the articles with the output $Y = \{0, 1\}$, where $Y = 0$ represents that the news is fake and $Y = 1$ represents that the news is true.

Each of the above branch modules and training steps will be described in detail in the following subsections.

3.1. Text feature extractor

Multimodal fake news detection techniques consider checking information of two main modalities: text and image. The text is the main expression of news events, which provides important clues for judging the credibility of news. Most methods use recurrent neural networks to model the input textual contextual information to capture the shallow features of the text, but the factual information that is extracted by such methods is very limited, and it is difficult to capture the semantic features of fake news. To better extract the contextual and semantic information from text, we extract text features mainly using pretrained bidirectional encoder representations from the transformer [BERT]. BERT is trained on large-scale datasets with powerful modeling capabilities, and it learns a large amount of common-sense information and semantic knowledge internally. In addition, BERT consists of stacked self-attention layers that better capture the connections between contexts.

Specifically, we encode the extracted text features of a list of words in each sentence as the input of the embedding vectors. The k-dimensional vector of the ith word is denoted as T_i, and the pretraining model that contains the 12-layer encoder bidirectional transformer is denoted as BERT. We input T into BRET to obtain a feature vector for each sentence as follows:

$$V_f = BRET\left(\left[T_f^0,\ T_f^1, \ldots, T_f^n\right]\right) \tag{1}$$

where V_f represents the feature vector of the fth sentence that is encoded by the BERT pretraining model and T_f^n is the k-dimensional feature vector of the nth word in the fth sentence. For each feature vector, we use the mean pooling operation to obtain the feature F_t of the whole text from all words; here, we determine that the text contains contextual and semantic information.

3.2. Visual feature extractor

The images in an article are also important for determining the authenticity of the article, and articles that contain images that do not match the text or are maliciously altered are often inauthentic. We start from two aspects: feature extraction from the image spatial domain and frequency domain information. The spatial domain information is used mainly for the semantic extraction of the image, and the frequency domain information is used mainly to detect whether the image has been modified or not. A modified image is more easily detected in the frequency domain space (Frank, et al., 2020).

In the spatial domain of an image. In recent works (Dosovitskiy, et al., 2020; Liu, et al., 2021), transformers have been widely and successfully applied to many image understanding tasks. In this paper, we employ SwinT, which will be pretrained in the ImageNet dataset, to extract the spatial semantic features of visual information. We use four Swin Transformer blocks to perform different levels of feature extraction on visual features.

Specifically, the image is first segmented into nonoverlapping patches by the patch segmentation module. Each patch is regarded as a mark. Here, we set the size of the patch to 4×4 and expand each RGB patch to obtain a $4 \times 4 \times 3$ feature vector, which is mapped to a feature space of DIM=96 by using the linear embedding layer. Next, a hierarchical representation is obtained and passed through 4 stages, where after each stage, each feature map is downsampled to half its previous size and the number of channels is doubled, and the result is input to the next stage. Here, it is expressed as:

$$Stage_i = SwinB(\sigma(W \times Stage_{i-1})) \tag{2}$$

where $Stage_i$ and $Stage_{i-1}$ represent the output and the input, respectively, of layer i and SwinB is a Swin Transformer block, which is composed of stacked self-attention networks. The layer and heads of self-attention that are contained in the 4-layer stage are set as [2, 2, 6, 2] and [3, 6, 12, 24], respectively. More details as shown Section 4.6. W is the learning parameter of downsampling. The eigenvectors that are output by the fourth layer are mapped to linear vectors through the linear layer. The obtained spatial domain features are denoted as F_s.

As shown in previous study (Frank, et al., 2020), the tampered images are more likely to be detected in the frequency domain space. To further explore the role of the frequency domain for fake news detection, we use the Discrete Fourier Transform (DFT) instead of the Discrete Cosine Transform (DCT) as a tool for the transformation from image space domain to frequency domain. Specifically, DFT converts the image from the spatial domain to the frequency domain represented by Fourier coefficients (complex-valued), where the imaginary part of the complex-valued corresponds to the phase component of the Fourier coefficients. The phase component contains the semantic features of the image that critically models the semantic similarity between the information of natural language and the authenticity of the image (Yang, Lao, Sundaramoorthi, & Soatto, 2020). First, the Discrete Fourier Transform (DFT) is applied to transform the spatial domain to the frequency domain of each image. To obtain a deeper feature, we use VGG19 as a feature extractor and input the imaginary and real parts of the frequency domain information into VGG19 after separating and concatenating them to obtain a deep semantic vector, as expressed in Formula (3):

$$F_f = VGG19(concat(IF_{imag}, IF_{real})) \tag{3}$$

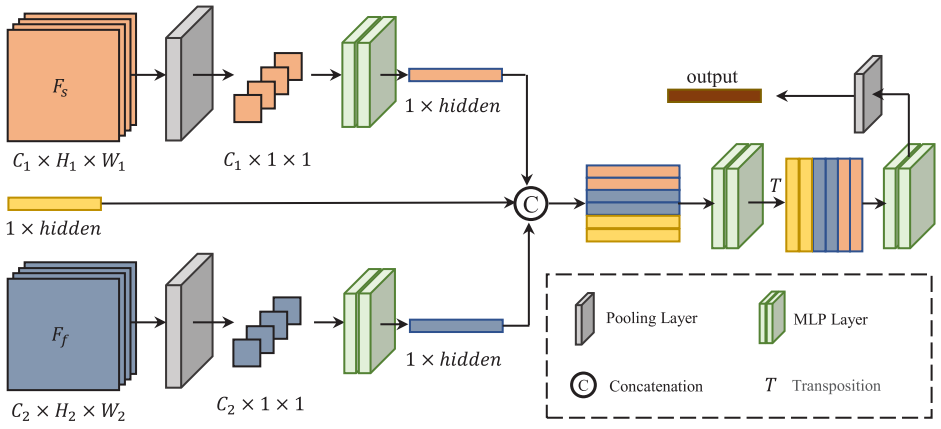

Fig. 3. Feature fusion process.

where IF_{imag} represents the imaginary part of the image frequency domain information and IF_{real} represents the real part of the image frequency domain information. After the discrete Fourier transform, the features contain more information compared to after the discrete cosine transform, which was used in the previous work.

3.3. Feature fusion

Image information and text information in news are complementary to each other, and we often compare images with text when reading news; hence, the fusion between text and image information is a crucial part of fake news detection. We design a progressive fusion approach for processing the shallow and textual information of images in stages to fully utilize the image and shallow information. We design the Mlp Mixer as a fusion module for images for fusing feature information between modalities at a finer granularity. The network structure of this module is illustrated in Fig. 3.

For images, the spatial domain feature extractor obtains features of different depths at different stage periods. According to the order of the extractors, we denote these 4 features of different depths as $Stage_{1,2,3,4}$. In the frequency-domain space, the feature map of the $2nd$, $4th$, $8th$, and $16th$ convolutional layers of VGG19 are denoted as V_1, V_2, V_3, and V_4, respectively. The text features that are extracted by the text feature extractor are denoted as T. Considering the $Stage_1$ fusion of shallow features as an example. First, the number of $Stage_1$ and V_1 channels C is expanded to 512 by a convolutional layer with a convolutional kernel of size 3. The feature map after the expansion is pooled on average to obtain a feature map of size b $\times$ 512 $\times$ 1 $\times$ 1. By means of linear mapping, the feature map is flattened and mapped to a 1000-dimensional feature vector. We expand three vectors of $Stage_1$, V_1, and T as (B, 3, 1000) dimensional vectors on dim1 to balance the distribution of modes, and these three feature vectors are stitched together in dimension 1 into a feature F with shape (B, 9, 1000). F is fused with features in dimension 2 using two layers of Mlp, and the fused features are transposed for the operation of feature fusion using Mlp, which realizes the fusion operation of the original features in dimension 1. Finally, inverse transformation is performed for feature compression, which recovers a feature vector of the same size as the feature T.

$$F_i = MLP\ Mixer(cat(Stage_i, V_i, T)) + T \tag{4}$$

where $MLP\ Mixer$ denotes the linear layer-based feature fusion, the concat operation is denoted as cat. We use ReLu and LayerNorm to improve the fusion capability. We combine the features that are obtained after fusion as a residual module and feature T to reduce the model risk and improve the feature extraction capability. The features of images and text are progressively fused from shallow to deep to improve the correlations between different modal features.

3.4. Fake news detector

We input the feature information after the fusion of multimodal representations into the fully connected layer, and the output of the fully connected layer is calculated as following softmax function to produce the distribution of categorical labels.

$$p = softmax(W_c X + b_c) \tag{5}$$

where W_c and b_c represent the parameters in the fully connected layer, where is used in the cross-entropy loss function.

$$L = -\sum [y^f log p^f + (1 - y^f) log(1 - p^f)] \tag{6}$$

where y^f represents the truthful value labeling the sample, 0 stands for a prediction of a fake news while 1 stands for an opposite prediction of a true news, and p^f denotes the probability that is predicted by the sample.

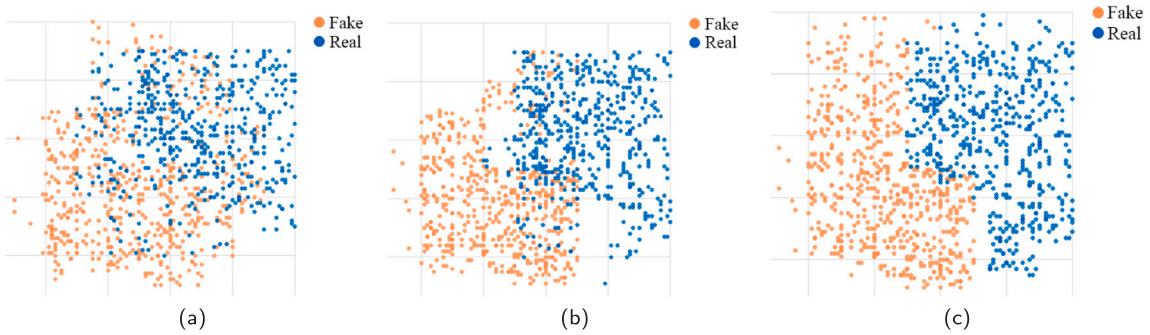

Fig. 4. Visualization of the performance of the feature representations that are learned by frequency-domain subnetworks, spatial-domain subnetworks, and MPFN on the Twitter dataset.

Table 1
Statistics of two real-world multimodal datasets.

Dataset	Label	Number	Total
Twitter	fake	7021	12995
	real	5974	
Weibo	fake	4749	9528
	real	4779	

4. Experiments

We validate our model on two datasets (Weibo and Twitter) in English and Chinese and evaluate the results quantitatively. A comparison between our proposed MPFN model and the state-of-the-art method is given in this section, including ablation experiments for validatity its performance. Specifically, a detailed training information and the parameter set is also provided.

4.1. Datasets

Twitter dataset: This dataset was collected and published by the MediaEval Benchmarking Initiative (Boididou, Papadopoulos, Kompatsiaris, Schifferes, & Newman, 2014) to evaluate multimodal performance. Specifically, this dataset include a sequential content of tweeted texts and images, composing of 6000 rumor posts and 5000 true posts as training set. The testing set varied different types of breaking news as many as 2000 posts. Posts contain only image or only text are not allowed to participate the training process or testing process.

Weibo dataset: This microblogging dataset has been widely applied in many recent studies for verifying the performance of multimodal fake news detection strategies, first collected by Jin, et al. (2017), only in Chinese articles or comments. The text content of the microblogging dataset consists mainly of Chinese text. Sina Weibo encourages its users to tip off any suspicious accounts and the malicious comments on heated discussion of current affairs. As a result, a non-profit committee consisting of reputable users manually check the cases and classifies them into fake and real news. Many recent studies on debunking system were applied as authoritative sources, and further classified and flagged by users between 2014 and 2016. Whether the news is real is depending on it is collectable from the authoritative Chinese news sources. In our work, the training set composing of 70% of the dataset, while the rest is further divided into validation set and test set in a ratio of 1:2. Table 1 demonstrates the statistical details of the two used dataset in this paper.

4.2. Experimental settings

In this paper, the F_1 value, Accuracy, and Recall of news prediction results are calculated, and these quantitative data are used as evaluation metrics for the models. There are 768 dimensionality of text feature are obtained from BERT in the text feature extractor. The size of the images are uniformly set to $224 \times 224 \times 3$ in the visual extractor, while the value of dimension of the final extracted feature vector is set to 1000. The learning rate and the batch size are set to 0.001 and 12, respectively. In the network, a leaky ReLu activating function is applied to all fully connected layers, while the value of dropout is 0.5 for the purpose of avoiding overfitting. Adam optimizer is also applied to better select the model parameters.

Table 2
Results of various methods on two datasets.

Dataset	Method	Accuracy	Fake news			Real news		
			Precision	Recall	F_1	Precision	Recall	F_1
Weibo	SVM-TS	0.640	0.741	0.573	0.646	0.651	0.798	0.711
	CNN	0.740	0.736	0.756	0.744	0.747	0.723	0.735
	GRU	0.702	0.671	0.794	0.727	0.747	0.609	0.671
	TextGCN	0.787	**0.975**	0.573	0.727	0.712	**0.985**	0.827
	EANN	0.782	0.827	0.697	0.756	0.752	0.863	0.804
	ATT-RNN	0.772	0.854	0.656	0.742	0.720	0.889	0.795
	MAVE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	SpotFake	0.869	0.877	0.859	0.868	0.861	0.879	0.870
	SpotFake+	**0.870**	0.887	0.849	0.868	0.855	0.892	0.873
	MPFN	0.838	0.857	**0.894**	**0.889**	**0.873**	0.863	**0.876**
Twitter	SVM-TS	0.529	0.488	0.497	0.496	0.565	0.556	0.561
	CNN	0.549	0.508	0.597	0.549	0.598	0.509	0.550
	GRU	0.634	0.581	0.812	0.677	0.758	0.502	0.604
	TextGCN	0.703	0.808	0.365	0.503	0.680	**0.939**	0.779
	EANN	0.648	0.810	0.498	0.617	0.584	0.759	0.660
	ATT-RNN	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	MAVE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	SpotFake	0.771	0.784	0.744	0.764	0.769	0.807	**0.787**
	SpotFake+	0.790	0.793	0.827	0.810	0.786	0.747	0.766
	MPFN	**0.833**	**0.846**	**0.921**	**0.880**	**0.809**	0.721	0.740

4.3. Baselines

In order to evaluate the performance of the proposed method, the prediction results on the test set of the above dataset is quantified and compared with the results of single-modal and multimodal fake news detection models.

Single-modal based approaches

- SVM-TS (Ma, Gao, Wei, Lu, & Wong, 2015): SVM-TS combines the advantages of heuristic rules and linear SVM classifiers for fake news detection.
- CNN (Yu, et al., 2017): CNN divides tweets into fixed-length sequences, and applies convolutional neural networks to achieve false information awareness. Its main contribution is accomplishing early detection tasks due to learning feature representations.
- GRU (Ma, et al., 2016): The GRU utilizes recurrent neural networks (RNNs) to learn hidden representations and can consider posts as variable time series using a multilayer GRU network.
- TextGCN (Yao, Mao, & Luo, 2019): In TextGCN algorithm, the entire corpus is modeled as a heterogeneous graph, and the words and posts are embedded by graph convolutional networks in the learning process.

Multimodal based approaches

- EANN (Wang, et al., 2018): EANN is a neural network from the literature that is based on an event adversarial mechanism. By introducing an event classifier as a secondary task, the model is guided to learn event-independent multimodal features. The model uses TextCNN and pretrained VGG19 for text and visual modal feature extraction, respectively, and splices the 2-modal features as multimodal feature expressions of fake news, which are input into a fake news classifier and a news event classifier.
- Att-RNN (Jin, et al., 2017): Att-RNN is a recurrent neural network from the literature that is based on an attention mechanism for fusing features of three modalities: text, visual and social contexts. Among them, the text part is modeled by LSTM, and the image part is modeled by pretrained VGG19 for feature extraction. For fairness of comparison, we remove the part that deals with social features in the concrete implementation.
- MVAE (Khattar et al., 2019): MVAE is a multitask model from the literature that combines a multimodal variational autoencoder and a fake news detector. Text and images are extracted by bidirectional LSTM and pretrained VGG19, respectively, and the spliced features of both are encoded as an intermediate expression for reconstructing input features and fake news classification.
- SpotFake (Singhal, Shah, Chakraborty, Kumaraguru, & Satoh, 2019): In the SpotFake algorithm, the pretrained language models, such as BERT, is applied to extract textual content, while VGG19 is set up to obtain image features.
- SpotFake+ (Singhal, et al., 2020): SpotFake+ is an upgraded version of SpotFake that uses pretrained XLNet (Yang, et al., 2019) models to extract text features.

4.4. Performance comparison

In the experimental section, Table 2 illustrates the detailed comparing results (see below in 5 aspects) between our proposed model MPFN and the baseline approaches on the two datasets.

1. As shown in Table 2, MPFN significantly outperforms the baseline methods in terms of most indicators on Weibo and Twitter datasets, thereby indicating that the progressive fusion approach that is proposed can effectively improve the performance of fake news detection.
2. In both datasets, the results of SVM-TS, which exemplifies the limitations of traditional methods, show that the performance of manual labeling is not sufficient for identifying fake news.
3. The deep learning methods CNN and RNN outperform SVM-TS, thereby reflecting the effectiveness of deep learning methods in fake news detection. In addition, TextGCN outperforms CNN and RNN on both datasets, which support the conclusion that the relationship between articles and words can be effectively captured through graph convolutional networks.
4. The multimodal model att-RNN considers the text-related part of each image due to its use of an attention mechanism; consequently, att-RNN outperforms GRU. In addition, the MVAE model outperforms approaches that consider only a single modality by using additional visual content as supplementary fake news detection method.
5. Performed results of both SpotFake and SpotFake+ indicate their effectiveness against other baselines on the Twitter and Weibo dataset. As a result, the pretrained BERT and XLNet can better enhance the model performance due to learned text features.

Our proposed MPFN performs best against other baselines on the Weibo and Twitter dataset. The results show that the model can effectively use the information between different levels to better capture the news representation for better fake news detection performance.

4.5. Ablation analysis

To verify the effectiveness of our proposed fusion strategy, fusion module, and frequency domain phase, we perform separate ablation analyses to evaluate the effectiveness of each component in improving the MPFN performance.

Effectiveness of fusion strategy. To verify the effectiveness of the MPFN fusion strategy, the model containing only the final fusion part is denoted as MPFN-B, which does not include the previous fusion process. Then, we split the fusion process sequentially, adding different stages of feature fusion sequentially from shallow to deep layers, and the obtained models with different structures are noted as MPFN-S (only include Stage1 fusion), MPFN-M (include Stage1 and Stage2 fusion), MPFN-L (include Stage1, Stage2, and Stage3 fusion) as well as the complete MPFN. Table 3 summarizes the experimental results of the different models on the Weibo and Twitter datasets. Compared with MPFN-B, MPFN-S improves the accuracy of fake news detection on the Weibo dataset by 0.9%, F1 by 2.3%, and recall by 1.1%. Further, we can also clearly observe that adding the remaining stages of fusion can improve the performance of our model to varying degrees. In addition, the performance gains from adding different levels of fusion are also seen in the Twitter dataset. Overall, fusing information from each modality at different levels is crucial for fake news detection. In particular, the inclusion of shallow-level feature fusion can improve the performance of the fake news detection task with better results.

Effectiveness of Fusion Module. To verify the effectiveness of our proposed fusion module, we retain the progressive fusion strategy and replace the Mlp Mixer with element-by-element addition (Addition), tensor splicing (Concat), and gated fusion (Gated), respectively. As shown in Fig. 5, the model using the Mlp Mixer fusion obtains optimal performance in terms of classification accuracy of real/fake news detection. Specifically, Mlp Mixer improved detection accuracy by 3.1% compared to Addition on the Weibo dataset, and improved detection precision by 3.1% and 2.6% for real/fake news, respectively. Compared with the gated fusion mechanism detection accuracy is improved by 1.7%, and the detection precision of real/fake news is improved by 1.8% and 0.9%, respectively. Similarly, our fusion module brings different levels of improvement to the Twitter dataset. The above results demonstrate that our proposed fusion module can effectively aggregate information under different modalities to improve the performance of the model.

Effectiveness of Frequency Domain. To enhance the role of semantic information in the frequency domain for fake news detection, we added phase information to explore the similarity with semantic information in natural language and the authenticity of the images. We prove the validity of the phase information by additional experiments on the phase component. As shown in Table 4, the performance of the model with the inclusion of phase information improves by 0.6% in the accuracy metric on the Weibo dataset. In summary, our inclusion of frequency domain phase information strengthens the model in terms of semantic guidance.

To verify the positive impact of DFT, we compare the performance of models using DFT and DCT on two commonly used benchmark datasets. As observed in Table 4, although the classification results obtained from the model applying DCT are slightly better than those obtained from the model applying DFT without phase in most metrics, the overall performance is generally lower when adopted to phase information scenarios. In conclusion, DFT outperforms DCT when applied in our model, and effectively improves the performance of fake news detection due to the detailed texture in amplitude and the semantic information in phase.

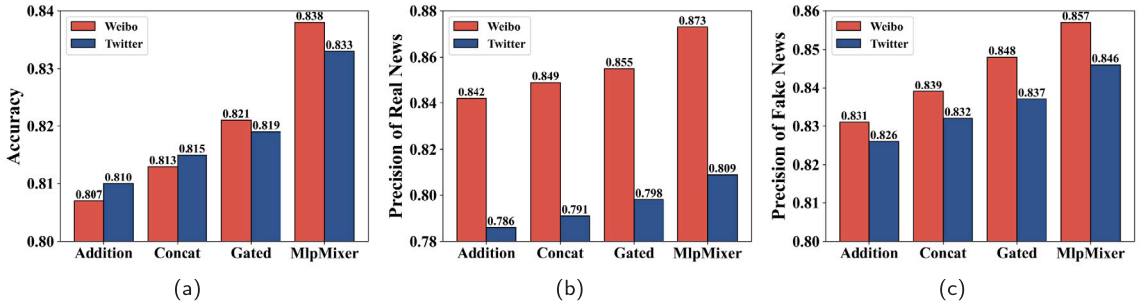

Fig. 5. Comparison of different fuser performance on two datasets.

Table 3
Experimental results of different fusion strategies.

Dataset	Method	Accuracy	Fake news			Real news		
			Precision	Recall	F_1	Precision	Recall	F_1
Weibo	MPFN-B	0.809	0.831	0.868	0.843	0.831	0.832	0.834
	MPFN-S	0.818	0.840	0.879	0.866	0.843	0.852	0.851
	MPFN-M	0.822	0.842	0.883	0.869	0.852	0.855	0.857
	MPFN-L	0.830	0.849	0.891	0.882	0.869	0.861	0.869
	MPFN	0.838	0.857	0.894	0.889	0.873	0.863	0.876
Twitter	MPFN-B	0.794	0.823	0.879	0.832	0.761	0.701	0.713
	MPFN-S	0.806	0.831	0.890	0.849	0.778	0.711	0.720
	MPFN-M	0.812	0.836	0.896	0.854	0.783	0.713	0.727
	MPFN-L	0.825	0.843	0.918	0.863	0.799	0.718	0.733
	MPFN	0.833	0.846	0.921	0.880	0.809	0.721	0.740

Table 4
Experimental results of spatial domain transforming frequency domain method.

Dataset	Method	Accuracy	Fake news			Real news		
			Precision	Recall	F_1	Precision	Recall	F_1
Weibo	DCT	0.834	0.851	0.889	0.882	0.869	0.860	0.873
	w/o Phase	0.832	0.849	0.884	0.881	0.867	0.854	0.871
	MPFN	0.838	0.857	0.894	0.889	0.873	0.863	0.876
Twitter	DCT	0.830	0.844	0.919	0.876	0.781	0.718	0.733
	w/o Phase	0.828	0.842	0.917	0.872	0.791	0.713	0.729
	MPFN	0.833	0.846	0.921	0.880	0.809	0.721	0.740

Table 5
Performance of parameter settings for different self-attention layers and heads.

Parameters	Layers [2,2,6,2] Heads [3,6,12,24]	Layers [2,2,18,2] Heads [3,6,12,24]	Layers [2,2,18,2] Heads [4,8,16,32]	Layers [2,2,18,2] Heads [6,12,24,48]
Weibo	0.838	0.828	0.835	0.825
Twitter	0.833	0.826	0.831	0.822

4.6. Parameters of self-attention

The Swin transformer encoder used in the image encoder senses the contextual information and spatial induction bias of the image, providing effective semantic and structural features for fake news detection. Inspired by Liu, et al. (2021), we set four parameters for the number of layers and heads of self-attention in each stage and tested the detection accuracy of the network with different parameters, and the results are shown in Table 5. From Table 5, we can observe that our method shows the best performance when the number of layers of self-attention in different stages are set to [2, 2, 6, 2] and the number of heads are set to [3, 6, 12, 24].

4.7. Visual data analysis

In our previous ablation experiments, we visually evaluated the effectiveness of each component in improving the performance of MPFN. Next, we visualize the feature representations that are learned by the frequency domain subnetwork (a), the spatial domain subnetwork (b), and MPFN (c), and we perform experiments on the Twitter dataset. The results are shown in Fig. 4, where we

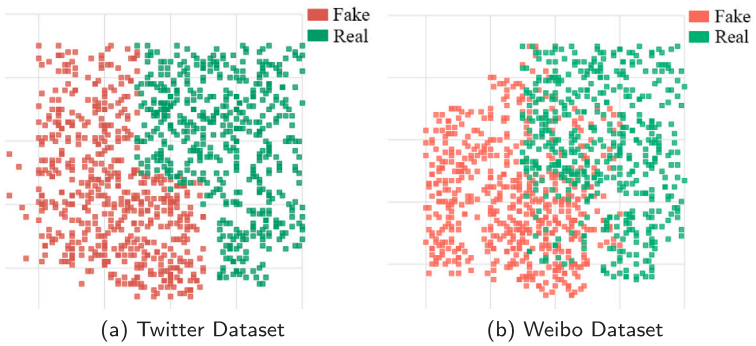

Fig. 6. Feature visualization performance of MPFN on two datasets.

can see the performance of each network in distinguishing whether the news is fake or not. We also observe that the divisibility of MPFN is significantly stronger than those of the other two networks, and the divisibility of the spatial domain network is better than that of the frequency domain network. Specifically, in the frequency domain subnetwork visualization graph, the overlap of feature representations is high because the images that were uploaded to social media were compressed, which makes the difference between the fake images that were originally compressed or tampered with and the real images in the frequency domain smaller. For the spatial domain subnetwork view, the feature representations are distinguishable, but some features are mostly overlap in the middle part. In contrast, the feature representations have relatively visible boundaries in MPFN's visual graphs. From the above observations, we conclude that the fake news can be better distinguished mainly due to features in the spatial domain, while less features in the frequency domain. Specifically, the spatial and frequency domains are complementary for the detected images; thus, fusing spatial and frequency domain information can lead to a better feature representation of MPFN and result in better performance than a single-domain network.

To analyze how MPFN distinguishes fake news from real news, Fig. 6 shows projections of our model's feature representations on the Twitter and Weibo datasets in a two-dimensional plane, and we observe that MPFN realizes satisfactory performance for both datasets. Due to the difference between Chinese and English text, the sequence of words obtained when segmenting Chinese text will be longer than the sequence of words segmented in English text, thereby resulting in a higher instance overlap rate in the Weibo dataset than in the Twitter dataset. We investigated the reasons for this. Stanford's Kumar team performed an experiment in which they hired markers and asked them to judge 320 pairs of real and fake news (Kumar, West, & Leskovec, 2016). The experiment yielded a 66% rate of people being able to correctly identify fake news and found that the longer the content of the fake news and the more markers there were, the easier it was to identify the news as real news. There are various communities in Chinese social media, and people in these communities share the same interests and characteristics. The echo chamber effect, which occurs on the Internet due to the presence of communities, causes people in the community to be confused by fake news. The echo chamber effect is the constant repetition and distortion of opinions in a limited environmental space. This causes people in the community to use longer sentences to describe an event; hence, the data in the Chinese dataset are longer, thereby rendering the overlap of instances in the Weibo dataset higher than that in the Twitter dataset.

5. Conclusions

In this paper, a multimodal fake news detection model with progressive fusion for determining the authenticity of misinformation is proposed. Specifically, this multimodal fake news model, named MPFN, reduces the negative effects of infodemic, and outperforms most previously established baseline methods by both considering the deep information and the shallow information of the images, which better fusing multimodal features with modalities interactions in detecting fake news. MPFN makes full use of various hierarchical features, fuses information from different modalities, and establishes fine-grained correlations between modalities during fusion. Experiments show that the MPFN model that is proposed in this paper outperforms other methods for fake detection on two public datasets.

Most quantitative studies on fake news are based on English-language data, and few Chinese-language related studies have been conducted, which is related to the difficulty of obtaining Chinese-language related data and the small amount of available labeled data. Therefore, in our next step research we will focus on the study of constructing a large corpus of Chinese fake news and detecting fake news by means of deceptive reviews and susceptible individuals in propagation process in posting-and-replying sequences.

CRediT authorship contribution statement

Jing Jing: Methodology, Conceptualization, Data curation, Formal analysis, Writing – original draft. **Hongchen Wu:** Project administration, Funding acquisition, Writing – review & editing. **Jie Sun:** Validation. **Xiaochang Fang:** Resources. **Huaxiang Zhang:** Supervision, Project administration.

Data availability

The data that has been used is confidential.

Acknowledgments

This research was funded by the National Natural Science Foundation of China under Grant 61702312, in part by the Key Research and Development Plan of Shandong Province under Grant 2019GGX101075, in part by the Natural Science Foundation of Shandong Province of China under Grant ZR2017BF019, and in part by the Taishan Scholar Project of Shandong, China under Grant ts20190924.

References

Boididou, C., Papadopoulos, S., Kompatsiaris, Y., Schifferes, S., & Newman, N. (2014). Challenges of computational verification in social multimedia. In *Proceedings of the 23rd international conference on world wide web* (pp. 743–748).

Bond, G. D., Holman, R. D., Eggert, J.-A. L., Speller, L. F., Garcia, O. N., Mejia, S. C., et al. (2017). 'Lyin'Ted','Crooked Hillary', and 'Deceptive Donald': Language of Lies in the 2016 US presidential debates. *Applied Cognitive Psychology*, *31*(6), 668–677.

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

Bovet, A., & Makse, H. A. (2019). Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications*, *10*(1), 1–14.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. In *Proceedings of the 20th international conference on world wide web* (pp. 675–684).

Chen, Y., Conroy, N. J., & Rubin, V. L. (2015). Misleading online content: recognizing clickbait as" false news". In *Proceedings of the 2015 ACM on workshop on multimodal deception detection* (pp. 15–19).

Chen, X., Zhou, F., Trajcevski, G., & Bonsangue, M. (0000). Multi-view learning with distinguishable feature fusion for rumor detection. *Knowledge-Based System*.

Chi, H., & Liao, B. (0000). A quantitative argumentation-based Automated eXplainable Decision System for fake news detection on social media. *Knowledge-Based System*.

Choi, Y., Uh, Y., Yoo, J., & Ha, J.-W. (2020). Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8188–8197).

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

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Feng, S., Banerjee, R., & Choi, Y. (2012). Syntactic stylometry for deception detection. In *Proceedings of the 50th annual meeting of the association for computational linguistics* (pp. 171–175).

Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., & Holz, T. (2020). Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning* (pp. 3247–3258). PMLR.

Giachanou, A., Zhang, G., & Rosso, P. (2020). Multimodal fake news detection with textual, visual and semantic information. In *International conference on text, speech, and dialogue* (pp. 30–38). Springer.

Jin, Z., Cao, J., Guo, H., Zhang, Y., & Luo, J. (2017). Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on multimedia* (pp. 795–816).

Jin, Z., Cao, J., Zhang, Y., Zhou, J., & Tian, Q. (2016). Novel visual and statistical image features for microblogs news verification. *IEEE Transactions on Multimedia*, *19*(3), 598–608.

Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019). Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference* (pp. 2915–2921).

Kumar, K. K., & Geethakumari, G. (2014). Detecting misinformation in online social networks using cognitive psychology. *Human-Centric Computing and Information Sciences*, *4*(1), 1–22.

Kumar, S., West, R., & Leskovec, J. (2016). Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on world wide web* (pp. 591–602).

Kumari, R., & Ekbal, A. (2021). Amfb: Attention based multimodal factorized bilinear pooling for multimodal fake news detection. *Expert Systems with Applications*.

Kwon, S., Cha, M., Jung, K., Chen, W., & Wang, Y. (2013). Prominent features of rumor propagation in online social media. In *2013 IEEE 13th international conference on data mining* (pp. 1103–1108). IEEE.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., et al. (2016). *Detecting rumors from microblogs with recurrent neural networks*. AAAI Press.

Ma, J., Gao, W., Wei, Z., Lu, Y., & Wong, K.-F. (2015). Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM international on conference on information and knowledge management* (pp. 1751–1754).

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2017). A stylometric inquiry into hyperpartisan and fake news. arXiv preprint arXiv:1702.05638.

Qi, P., Cao, J., Yang, T., Guo, J., & Li, J. (2019). Exploiting multi-domain visual information for fake news detection. In *2019 IEEE international conference on data mining* (pp. 518–527). IEEE.

Satu, M. S., Khan, M. I., Mahmud, M., Uddin, S., Summers, M. A., Quinn, J. M., et al. (2021). TClustVID: A novel machine learning classification model to investigate topics and sentiment in COVID-19 tweets. *Knowledge-Based Systems*.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Singhal, S., Kabra, A., Sharma, M., Shah, R. R., Chakraborty, T., & Kumaraguru, P. (2020). Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). In *Proceedings of the AAAI conference on artificial intelligence. Vol. 34* (10), (pp. 13915–13916).

Singhal, S., Shah, R. R., Chakraborty, T., Kumaraguru, P., & Satoh, S. (2019). Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data* (pp. 39–47). IEEE.

Song, C., Ning, N., Zhang, Y., & Wu, B. (2021a). Knowledge augmented transformer for adversarial multidomain multiclassification multimodal fake news detection. *Neurocomputing*.

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

Tran, N.-T., Tran, V.-H., Nguyen, N.-B., Yang, L., & Cheung, N.-M. (2019). Self-supervised gan: Analysis and improvement with multi-class minimax game. arXiv preprint arXiv:1911.06997.

Tuan, N. M. D., & Minh, P. Q. N. (2021). Multimodal fusion with BERT and attention mechanism for fake news detection. Preprint arXiv:2104.11476.

Volkova, S., Shaffer, K., Jang, J. Y., & Hodas, N. (2017). Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th annual meeting of the association for computational linguistics* (pp. 647–653).

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., et al. (2018). Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th Acm sigkdd international conference on knowledge discovery & data mining* (pp. 849–857).

Wang, Y., Ma, F., Wang, H., Jha, K., & Gao, J. (2021). Multimodal emergent fake news detection via meta neural process networks. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining* (pp. 3708–3716).

Wu, K., Yang, S., & Zhu, K. Q. (2015). False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st international conference on data engineering* (pp. 651–662). IEEE.

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

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

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, *32*.

Yang, Y., Lao, D., Sundaramoorthi, G., & Soatto, S. (2020). Phase consistent ecological domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9011–9020).

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

Yao, L., Mao, C., & Luo, Y. (2019). Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence. Vol. 33* (01), (pp. 7370–7377).

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

Zhang, G., Giachanou, A., & Rosso, P. (2022). SceneFND: Multimodal fake news detection by modelling scene context information. *Journal of Information Science*, Article 01655515221087683.

Zhou, X., Zafarani, R., Shu, K., & Liu, H. (2019). Fake news: Fundamental theories, detection strategies and challenges. In *Proceedings of the twelfth ACM international conference on web search and data mining* (pp. 836–837).