frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 21 April 2023
DOI 10.3389/fcomp.2023.1159063

OPEN ACCESS

EDITED BY
Gaolei Li,
Shanghai Jiao Tong University, China

REVIEWED BY
Chong Di,
Qilu University of Technology, China
Xiao Yang,
Shanghai Jiao Tong University, China
Yuchen Liu,
North Carolina State University, United States

*CORRESPONDENCE
Ying Guo
✉ guoying@ncut.edu.cn

SPECIALTY SECTION
This article was submitted to Computer Security, a section of the journal Frontiers in Computer Science

RECEIVED 05 February 2023
ACCEPTED 17 March 2023
PUBLISHED 21 April 2023

CITATION
Guo Y, Ge H and Li J (2023) A two-branch multimodal fake news detection model based on multimodal bilinear pooling and attention mechanism. *Front. Comput. Sci.* 5:1159063. doi: 10.3389/fcomp.2023.1159063

A two-branch multimodal fake news detection model based on multimodal bilinear pooling and attention mechanism

Ying Guo*, Hong Ge and Jinhong Li

Department of Computer Science, North China University of Technology, Beijing, China

Introduction: Fake news spread in various areas has a major negative impact on social life. Meanwhile, fake news with text and visual content is more compelling than text-only content and quickly spreads across social media. Therefore, detecting fake news is a pressing task for the current society.

Methods: Concern the problem of extracting insufficient features, and the inability to merge multi-modality features effectively in detecting fake news. In this article, we propose a method for detecting fake news by fusing text and visual data. Firstly, we use two-branch to learn hidden layer information of modality to obtain more helpful features. Then we proposed a multimodal bilinear pooling mechanism to better merge textual and visual features and an attention mechanism to capture multimodal internal relationships for the detection of fake news.

Results and discussion: The experimental results demonstrated that our methodology outperformed the current state-of-the-art methodology on publicly accessible Weibo and Twitter datasets.

KEYWORDS

fake news detection, multimodal data fusion, multimodal bilinear pooling, self-attention, two-branch's network

1. Introduction

Fake news refers to fake information intentionally created for political or economic purposes and is characterized by its rapid spread. The proliferation of fake news not only triggers a storm of public opinion but also manipulates public events, causing more direct harm to society (Meel and Vishwakarma, 2020). The wide application of social media like social networking sites and instant messaging has made it easier for manipulators of public events to make up or change facts. Microblogs and Twitter, for example, encourage users to create their own content and publish, share, communicate, and spread it through social networking platforms, which makes it more challenging to control fake news (Shu et al., 2017). During the global fight against coronavirus, a plethora of fake news stories involving conspiracies and vested interests, such as the miracle drug Double Yellow Lotus Oral Liquid and the 5G spreading virus, were deliberately fabricated and spread on social media, causing many unnecessary mass panic incidents and farces in international relations (Salvi et al., 2021). How to stop the collateral damage caused by the manipulation of public events through social media platforms has become a social issue worth exploring during the current outbreak of coronavirus. Reliable methods and technologies are urgently required to stop bad epidemic prevention and control culture, screen social media fake news, and preserve a positive environment for the dissemination of accurate information.

As the media environment has changed, the public's access to news information has changed to online news and social platforms, and the structural form of fake news has also changed with the media ecosystem, from text to rich images and videos (Wu et al., 2015). Heterogeneity and multimodality make fake news carry richer and more intuitive information, attracting public attention and spreading (Jin et al., 2016). Therefore, deeper text and image content mining in fake news is crucial for fake news detection.

Early detection research suggested using single-modal text linguistics and content to identify fake news (Castillo et al., 2011; Qazvinian et al., 2011; Ma et al., 2016, 2019). It might be difficult to recognize fake news by relying solely on the claims made in the text because these claims are frequently written with the intent of misleading readers. Because of the proliferation of social media platforms, researchers detecting fake news are increasingly turning to the use of image-based information (Qi et al., 2019). Text and image information provides various and complementary information due to the multi-modality of fake news content. As a result, recent research has gradually shifted from single-modal to multi-modal to detect fake news (Jin et al., 2017; Wang et al., 2018; Khattar et al., 2019; Singhal et al., 2019; Zhang et al., 2019; Zhang T. et al., 2020).

However, the existing methods have certain shortcomings. For the feature extraction process, the current research does not take into account the hidden information of features when extracting features from news texts and images. For the feature fusion process, text information and image information are simply concatenated, without considering effective fusion between modalities. Therefore, the research questions of this study can be summarized as the following two points:

- Feature Extraction Problem

  In the task of fake news detection, text modality and image modality are involved, so feature extraction is needed for both modalities separately. Due to the differences in the presentation and description of text and image, how to extract the effective features of different modalities from the news content is a research question. Second, there is a wide variety of fake news published every day. Most existing work tends to learn the features of specific news events and cannot be converted to unseen news events. How to propose effective features from the newly emerged news events also is a question.
- Feature Fusion Problem

  At the present stage, most of the fake news detection methods using textual features combined with visual features only complete the representation of fused features by simple concatenation of multiple modal feature vectors, which is not sufficient to express the complementarity and difference between multimodal data. It is a major research problem to ensure that the modal features can complement each other's different features.

In summary, to address the problems of feature extraction, we propose using a two-branch network for deep and shallow feature extraction under the premise of using a deep pre-training model to obtain feature vectors at different levels. We also use a domain adversarial network for adversarial training to obtain common features in different event domains to solve the problem of the generalizability of the model. To address the feature fusion problem, we will propose a multimodal fusion method based on multiple fusion mechanisms. For inter-modal content, a multimodal bilinear pooling method is used to fully combine the unique dimensional information of each position of text and image, and for intra-modal content of each modality, a self-attention mechanism is used to enhance the self-content so as to maintain the integrity and diversity of features.

So we present a two-branch multimodal fake news detection model based on multimodal bilinear pooling and attention mechanism (MBPAM). The model has four parts: multimodal feature extractor module, multimodal feature fusion module, domain adversarial module, and fake news detector module. The multimodal feature extractor module contains a text extractor and an image extractor. The text extractor uses a pre-trained BERT to extract sentence and word features that have contextual meaning from two branches, while the image content extractor uses a residual network. ResNet is used to collect image features from two branches. The multimodal feature fusion module includes an inter-modal feature fusion module based on multimodal bilinear pooling, as well as an intra-modal information enhancement module based on self-attention mechanism to accomplish effective information interaction. Additionally, the domain adversarial module is employed to increase the generality of domain features and eliminate features from specific news domains. In the final step, the features entered into a fake news detector to be detected. The most important contribution made by this study can be summed up as follows:

- The state-of-the-art of model Bert in NLP and model Resnet in CV are employed in this study. Bert replaces the past LSTM to extract the textual model, while Resnet is the replacement of VGG. The latest models may exert the most comprehensive roles in feature extraction.
- We employ two branches to extract useful features from the text and image modalities' hidden layer information: one branch concatenates and fuses shallow features, and the second branch fuses more complex forms to mine deeper relevance.
- We propose a multimodal bilinear pooling method to better fuse information between inter-modalities and self-attention mechanism to enhance intra-modal information. Thus, we exploit the interaction inside and outside the heterogeneous modal data to improve the performance of news detection.
- Experiments on two publicly available datasets show that our model performs better than state-of-the-art methods for detecting fake news.

2. Related work

2.1. Fake news detection

Fake news refers to information that has been fabricated intentionally and whose authenticity can be proven false (Shu et al.,

2017). Two main categories of fake news detecting methods are now in use: single-modal approaches and multimodal approaches.

2.1.1. Single-modal approaches

Text-based approaches and image-based approaches are the two types that are included in the category of single-modal approaches.

Text-based approaches mainly used linguistics and textual content. Castillo et al. (2011) detected fake news by textual counting information, such as special characters, and the number of links. Qazvinian et al. (2011) used Bayesian networks as classifiers to distinguish fake news by analyzing the text's topic features. These features, however, are human-designed and may pose complex problems, introducing prejudices and design difficulty. Thus, deep learning technologies afterward address this issue. To better understand how the text is represented, Ma et al. (2016) introduced recurrent neural networks. Ma et al. (2019) presented a way based on generative adversarial networks that can obtain low-frequency but more discriminative features through adversarial training.

In addition to the text features, images have been shown to play a pivotal role in disinformation detection (Wu et al., 2015; Jin et al., 2016). Earlier studies mainly used basic statistical features of images, such as size and radio, which failed to fully extract the semantic information. Recent studies usually use pre-trained deep CNN networks to extract image features. Qi et al. (2019) fused information from the frequency and pixel domains of images for fake news identification using an attention mechanism.

Even though fake news detection has improved somewhat from a single-modal standpoint, the information utilization and detection performance are low when the problem is studied only from one perspective.

2.1.2. Multimodal approaches

Multimodal approaches use text modal information and image modal information to detect fake news. Jin et al. (2017) proposes using recurrent neural networks to detect fake news, and the neural network has an attention mechanism that would allow the network to easily merge text and image features. To exclude the interference of specific events, Wang et al. (2018) suggested a method for identifying fake news that employs event adversarial neural networks that find commonalities between different events. Zhang et al. (2019) used event memory networks to capture potential topic information unrelated to specific events and obtained better generalization ability for emerging events. Khattar et al. (2019) suggested using a multimodal variable auto-encoder that integrates a variable score auto-encoder with a classifier for disinformation detection to learn common latent representations across modalities. Zhang T. et al. (2020) presented a model that pinpoints unusual pieces of information by using domain classifiers. These classifiers map the features of a variety of events to the same space. Singhal et al. (2019) achieved successful classification by utilizing VGG19 to capture image features and BERT to extract text features. These two sets of extracted features were concatenated together as a joint representation for the purpose of classification.

Despite the advancement of multimodal fake news detection, most existing techniques do not fully exploit the link between modalities. In particular, they only do simple concatenate operations in the fusion stage. Due to the differences between text modality and image modality, completing the representation of fused features by simple concatenation between multiple modal feature vectors is not sufficient to express the complementarity and differences between multimodal data, which will lead to a biased performance in the final detection task. Therefore, based on existing research, in order to get around the shortcomings of the previous work, this paper develops a novel feature fusion technique.

2.2. Multimodal data fusion

With the development of technology, information exists in various forms, and different forms of existence or sources of information can be called as one modality. Thus, those data consisting of two or more modalities are called multimodal data (Gao et al., 2020). Multimodal data fusion is responsible for effectively integrating information from multiple modalities and drawing the advantages of different modalities. The information obtained from different modal forms naturally differs, so the fusion of information from different modalities allows the establishment of complementary relationships between modalities to eliminate redundant features and make the features more representative (Lahat et al., 2015). Text and images are the two modalities that are utilized in fake news, and the utilization of multimodal fusion of these two modalities is required. Three major techniques for fusing text and image features are utilized in deep learning: simple operation-based fusion, attention-based fusion, and bilinear pooling-based fusion (Zhang C. et al., 2020). All three methods achieve information fusion by correlating the feature vectors thus making the feature information more representative.

2.2.1. Simple operation-based

Deep learning can use simple operations like concatenation or weighted summation to combine vectorized features from different sources of data. Because deep models can be trained together, the high-level feature extraction hierarchy can be changed to fit the operations. This means that these operations usually have few or no correlation parameters.

2.2.2. Attention-based

The process of fusion typically makes use of attention processes. An "attention mechanism" refers to the weighted sum of a set of vectors with scalar weights, which are dynamically generated by a small "attention" model at each time step. This is what is typically meant when people talk about "attention mechanisms" (Bahdanau et al., 2014; Graves et al., 2014). Multiple outputs are often used to generate multiple sets of dynamic weights for summation, which can preserve additional information by concatenating the results derived from each glimpse.

2.2.3. Bilinear pooling-based

To merge text feature vectors and image feature vectors into a common representation space, the method called bilinear pooling is widely used. This is accomplished by computing the outer product of both vectors, which makes it possible for multiplicative interactions to take place among all of the elements in both vectors (Tenenbaum and Freeman, 2000). Meaning this method is more expressive.

For news, images usually show specific information about news events with visual effects, while the corresponding text descriptions describe news events in language. Whether it is text or image information, it will bring different effects. Text features are usually logically developed in a linear description, while images are more spatially described. If the difference between the two features is only through ordinary concatenate, there will be redundancy or lack of information. A multimodal bilinear pooling approach that relies on the outer product of feature vectors was used to provide a better fusion effect. To address the issue of the difference between the two modalities of text and image, this method fully integrates the text and image features of each dimension at a position. At the same time, some important words in the text modality and regions in the image modality will provide more information, so these textual and image features should be assigned greater weights. So, to prevent the loss of information after fusion, it should be strengthened in its own modality by means of self-attention. Thus, we propose to use multimodal bilinear pooling and self-attention mechanisms to obtain fusion feature representation. Therefore, a new fake news detection model of fusion feature representation is constructed.

3. Methodology

This part outlines a multimodal approach to identifying fake news. The model consists of four parts. The first part is the multimodal feature extractor module, where we use two branches to extract both textual and visual features. Text Branch-1 and Text Branch-2 represent extracting different levels of text features, and similarly, Image Branch-1 and Image Branch-2 represent extracting different levels of image features, which will be explained later. The second part is the multimodal feature fusion module, including both the inter-modal fusion module and intra-modal fusion module. Among them, the inter-modal information is fused using a multimodal bilinear pooling module, and the intra-modal information is augmented using a self-attention module. The third part is the domain adversarial module, which eliminates specific event-related features. The final part is the fake news detector module, by which the features are fed into to detect classification. Figure 1 shows the architecture of our proposed model.

3.1. Multimodal feature extractor module

Based on the content of news, multimodal feature extractor modules are divided into two categories: text feature extractor and image feature extractor.

3.1.1. Text feature extractor

This study uses BERT to get the underlying semantics of textual content more effectively (Devlin et al., 2018). BERT is a bi-directional pre-trained language model based on transformer with powerful semantic information modeling capability to extract deep contextual information. It has been demonstrated that the last four hidden levels of BERT contain rich information. The hidden layers of BERT can record different types of information features (Horne et al., 2020). As a result, BERT is used to obtain different levels of information within context, followed by some operations with two branches, as is shown in Figure 2. Text Branch-1 uses the pooling layer output of BERT as a sentence vector with contextual meaning, adding a fully connected layer to resize the shape. The first branch's output is denoted as T_f. We concatenate the final four hidden layers as contextual token embedding in Text Branch-2. The next step is to obtain more features from the different sets of word vectors by convolutional filters of sizes 2, 3, 4, and 5 in 1D-CNN layers. The embedding vectors are run through 1D-CNN layers, and the outputs are stacked to produce multi-granularity word features T_m for the next feature fusion.

3.1.2. Image feature extractor

To extract features from images and acquire their deep semantic features, deep CNN networks are used. However, there are degradation and gradient diffusion issues with deep networks (Bengio et al., 1994). ResNet mitigates the gradient problem by spanning the input across layers and adding it to the result of convolution through the shortcut connections. Thus, ResNet performs nicely in the fields of detection, segmentation, and recognition (He et al., 2016). As a result, a pre-trained ResNet-50 model is employed in this study to extract features from visual news information. A similar two-branch structure is used to generate features in different directions of the image, as is shown in Figure 3. The output of ResNet-50's last pooling layer is extracted by Image Branch-1, This then passes it through two fully connected layers to adjust the dimension. This branch's ultimate visual representation is defined as V_f. Image Branch-2 takes the 3D tensor representation of the output of the final convolutional layer module and transforms it into a 2D tensor. It is then used as input into a fully connected layer to create the image feature V_m for the following feature fusion.

3.2. Multimodal feature fusion module

To acquire the relationship between different modalities and obtain important information within modalities, an inter-modal feature combination module and an intra-modal feature enhancement module are constructed respectively.

3.2.1. Inter-modal feature combination module

For a more effective fusion of text and image features, it is necessary to take advantage of both types of data. This study adopts the multimodal bilinear pooling method to fully combine each dimension information of each position of text and image. Multimodal bilinear pooling comes from bilinear pooling (Lin et al., 2015) and was originally used in the VQA domain (Antol

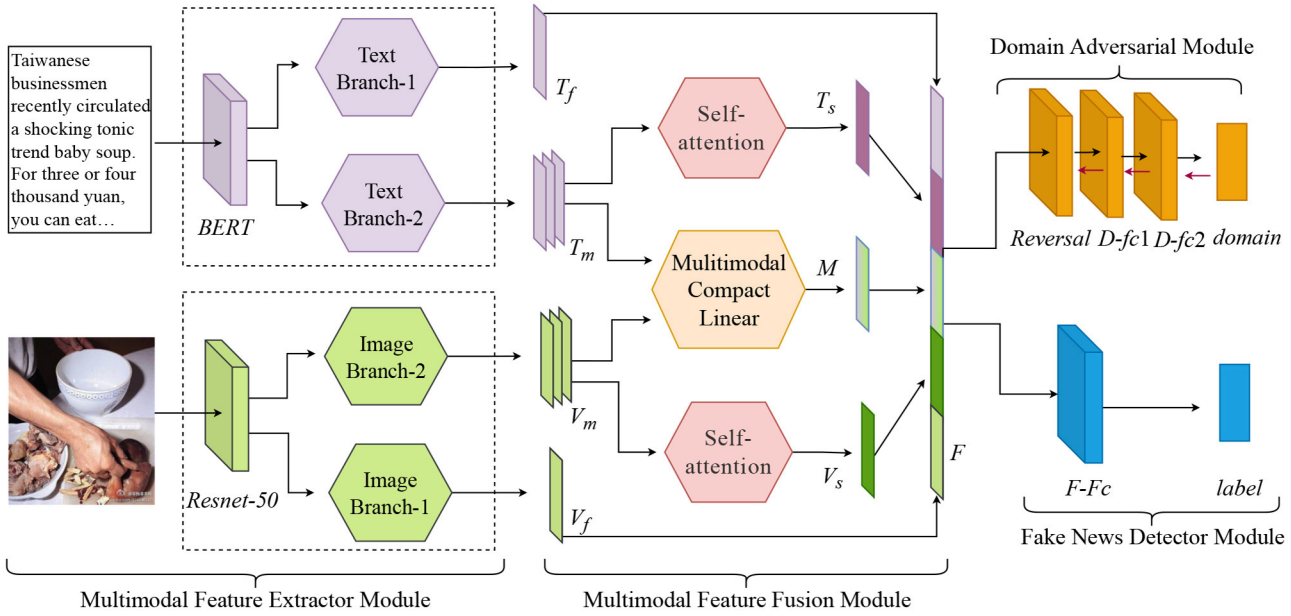

FIGURE 1
Network architecture of the proposed model based on Multimodal Bilinear Pooling and Attention Mechanism (MBPAM). It has four components: multimodal feature extractor module, multimodal feature fusion module, domain adversarial module, and fake news detector module.

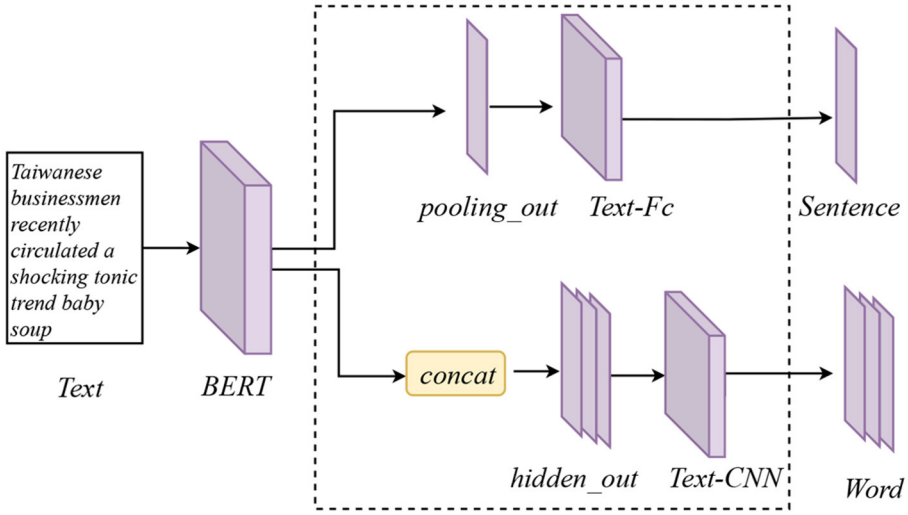

FIGURE 2
Text feature extractor module. It contains two branches, the upper branch extracts the entire vector semantics of the sentence (T_f in Figure 1) and the lower branch is used to extract the word vector semantics (T_m in Figure 1).

et al., 2015). Its essence is to calculate the outer product of two features, which has been indicated as a fabulous tool in data fusion.

The bilinear pooling method, on the other hand, allows all of the components of the two vectors to engage in a multiplicative way of interaction with one another. This directly leads to the resulting feature dimension being too high and the calculation being very complicated. Therefore, we need a way that projects the outer product to a lower dimensional space and avoids directly computing the outer product. The Count Sketch (Charikar et al., 2002) strategy can project the outer product into a lower dimensional space and can be used to solve the issue.

The specific dimensionality reduction process pseudo-code of Count Sketch is shown in Algorithm 1. The procedure is as follows: first, vectors $s \in \{-1, 1\}^n$ and $h \in \{1, \ldots, d\}^n$ are initialized for use in the dimensionality reduction process. For the elements $v[i]$ that need to be dimensioned down, use $h \in \{1, \ldots, d\}^n$ to find the target index $j = h[i]$ of the y vector, and add $s[i] \cdot v[i]$ to $y[j]$, which projects a vector from $v \in R^n$ to $y \in R^d$. At the same time, the outer product of two vectors does not need to be calculated

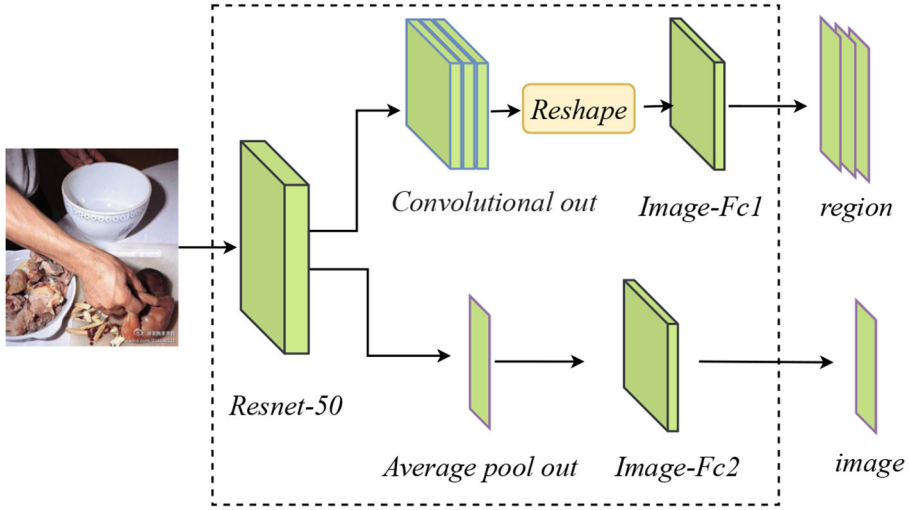

FIGURE 3
Image feature extractor module. It contains two branches, the upper branch is used to extract regional information (V_m in Figure 1) for multimodal fusion and the lower branch is used to extract overall deep image information (V_f in Figure 1).

```
input: v ∈ R^n
output: y ∈ R^d
if h,s not initialized then
    for i ← 1...n do
        sample h[i] from {1,...,d}
        sample s[i] from {-1,1}
v = Ψ(v,h,s,n)
procedure Ψ(v,h,s,n)
    y = [0,...,0]
    for i ← 1...n do
        y[h[i]] = y[h[i]] + s[i] · v[i]
    return y
```

Algorithm 1. Count sketch.

directly because Pham and Pagh (Pham and Pagh, 2013) showed that the count sketch of the outer product can be represented as a convolution of both count sketches, which avoids the need to directly calculate the outer product. The calculation formula is as follows:

$$\Psi(x \otimes q, h, s) = \Psi(x, h, s) * \Psi(q, h, s) \tag{1}$$

where Ψ represents the Count Sketch method and $*$ is the convolution operator.

Under the analysis, two features are mapped to low-dimensional space for convolution calculation. As far as we are concerned, based on the convolution theorem, the element-wise product in the frequency domain is equal to convolution in the time domain. So, the two vectors are transformed to the frequency domain through Fourier transform, and the vector product is made in the frequency domain, then the result is inverse Fourier transformed to the time domain space to get the final result. The formula is as follows:

$$x' = \Psi(x, h, s) \tag{2}$$

$$q' = \Psi(q, h, s) \tag{3}$$

$$x' * q' = FFT^{-1}(FFT(x') \odot FFT(q')) \tag{4}$$

where $\odot$ refers to element-wise product.

The aforementioned process is the execution flow of the Multimodal Compact Bilinear module (MCBP) (Fukui et al., 2016). The MCBP module is shown in Figure 4. Additionally, the inter-modal fusion module is built using the MCBP module.

Therefore, The MCBP module merges the text feature T_m and the image feature V_m into a multimodal feature. First, the Count Sketch method is used to reduce the dimensions of T_m and V_m, respectively. Following the Fast Fourier transform, the two features are element-wise products, and then the inverse Fast Fourier transform is used to produce the fusion feature M. The equation is as follows:

$$M = MCBP([T_m, V_m]) \tag{5}$$

3.2.2. Intra-modal feature enhancement module

The self-attention mechanism is able to evaluate the significance of various aspects of the information based on the weight of the information, enhancing the significance of the essential information while diminishing the significance of the superfluous information (Vaswani et al., 2017). The attention mechanism can weigh the inputs differently depending on their different positions or the importance of the features. The attention mechanism usually consists of three components: a query vector, a key vector, and a value vector. The attention value is obtained by calculating the attention scores of the given query vector and the corresponding weights, and then

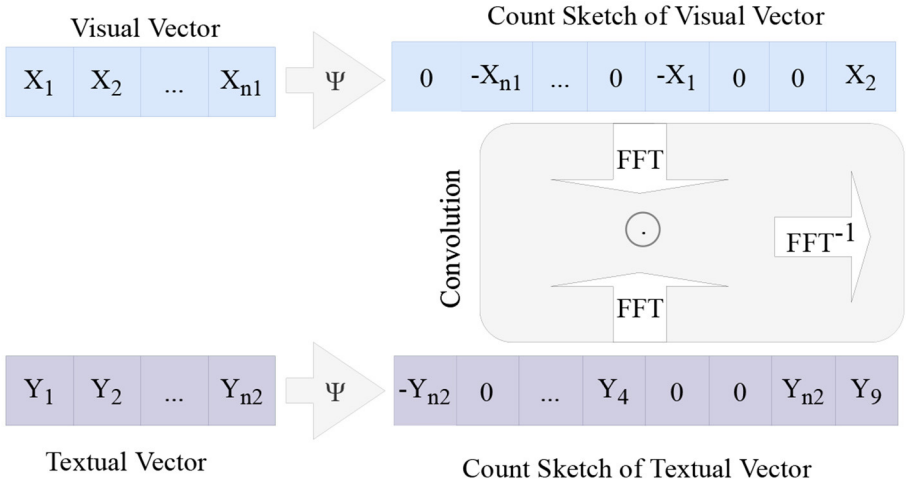

FIGURE 4
Multimodal compact bilinear pooling module. The image features and text features are downscaled by Count Sketch, then the dot product operation is performed by Fast Fourier Transform to the frequency domain, and finally, the fused features are obtained by inverse Fast Fourier Transform.

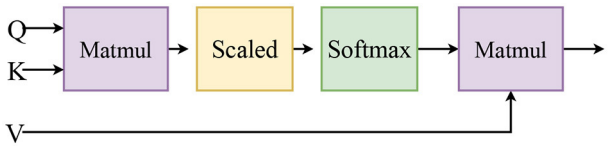

FIGURE 5
The scaled Dot-Production attention module. First, calculate the attention scores of *Q* and *K*, then use *softmax* to normalize the calculated attention weight scores, and finally weight the scores and *V* to obtain the final attention value.

the attention scores are used as the weighting coefficients for the weighting sum to obtain the final Attention Value. It is currently being utilized in a substantial capacity within the domain of text processing. Therefore, to prevent the loss of information after fusion, it should be strengthened from its own modality. Thus, the intra-modal information enhancement module is implemented by using a self-attention mechanism. It can enhance the information of the image modality and the text modality itself, thereby maintaining the integrity and diversity of features. Scaled Dot-Production attention is employed to implement the self-attention mechanism. The flow is shown in Figure 5.

The feature matrices Q, K, and V stand for query, key, and value. Q, K, and V are in the feature enhancement module from the same mode of operation. The first step is to compute the weight coefficients based on Query and Key, and the second step is to weigh the sum of value based on the weight coefficients. Calculating the difference between feature matrices yields self-modality key information.

First, the feature enhancement is performed for the text word features T_m to find the key features. Q, K, and V are all text words features. And the calculation process is as follows:

$$Q = W_{qt}T_m, K = W_{kt}T_m, V = W_{vt}T_m \tag{6}$$

$$T_s = softmax(\frac{Q \times K^T}{\sqrt{d}}) \times V \tag{7}$$

where T_s is the feature of the text words enhanced, W_{qt}, W_{kt}, and W_{vt} is the weight matrix, and d is the dimension of the input text word feature.

Second, the image features V_m are enhanced to find the key image information. Q, K, and V are the modal features of the image. The calculation process is shown as follows:

$$Q = W_{qv}V_m, K = W_{kv}V_m, V = W_{vv}V_m \tag{8}$$

$$V_s = softmax(\frac{Q \times K^T}{\sqrt{d}}) \times V \tag{9}$$

where V_s is the feature of the image itself enhanced, W_{qv}, W_{kv}, and W_{vv} is the weight matrix, and d is the dimension of the input image feature.

Finally, the text features T_f and T_s, the image feature V_f and V_s, and the fusion feature M concatenated to get the final feature F.

$$F = [T_f, T_s, V_f, V_s, M] \tag{10}$$

3.3. Domain adversarial module

A multimodal feature is derived based on the previous work. If the features are directly fed into the fake news detector, only the domain news contained in the training set will be detected. Inspired by Ganin and Lempitsky (2015) and Wang et al. (2018), we add a domain adversarial to improve the model's generalizability. Using the input of features F, it can be assigned to one of the K domain categories. Among them, multimodal fusion module map data

from different domains into the same feature space and attempt to trick the domain adversarial to increase the discriminative loss, thus getting the common features. On the other hand, the domain adversarial attempts to identify news by locating information about a specific event contained in the fused features. To compute the loss, the domain adversarial makes use of the cross-entropy. The loss function is defined as follows:

$$Loss_d = -E_{(x,y_e)\sim(X,Y_e)}[\sum_{k=1}^{K} y_e log(D(F;\theta_d))] \quad (11)$$

where D is the domain adversarial, θ_d is the parameter, and $D(F;\theta_d)$ is the predicted domain category probability. Y_e represents the event domain set, X is the news set, and y_e refers to the actual domain category classification of news x.

3.4. Fake news detector module

The fake news detector determines the truth of news by inputting the multimodal feature F, and the process is as follows:

$$\hat{y} = G(F;\theta_g) \quad (12)$$

where G is the fake news detector, θ_g is the parameter, and $\hat{y}$ is the predicted probability to be true and fake classification.

Cross-entropy is employed as the loss function for the fake detector. The loss of the fake detector module needs to be kept to a minimum. The process is as follows:

$$Loss_g = -E_{(x,y)\sim(X,Y)}(ylog(\hat{y}) + (1-y)log(1-\hat{y})) \quad (13)$$

where Y is the set of news true and fake label categories, X is the news set, and y refers to the actual label of news x.

A fake news detector is employed during model training to increase the effectiveness of identifying fake news, and a domain adversarial is used to obtain generic features. The model's objective is to discover the ideal parameters by minimizing the loss function. A gradient inverse layer is used to create an adversarial relationship between the domain adversarial and the feature fusion (Ganin and Lempitsky, 2015). So, the total loss is defined as follows:

$$Loss(\theta_g,\theta_d) = Loss_g(\theta_g) - \lambda Loss_d(\theta_d) \quad (14)$$

where λ is a coefficient to balance the loss of fake news detector and domain adversarial.

4. Experiments

4.1. Dataset

Two publicly accessible datasets, the Weibo dataset and the Twitter dataset, are used for our model training.

TABLE 1 The statistics of the real-world datasets includes Weibo dataset and Twitter dataset.

Dataset	Label	Number	All
Weibo	Fake	4,749	9,528
	Real	4,779	
Twitter	Fake	7021	12,995
	Real	5,924	

4.1.1. Weibo dataset

Weibo dataset (Jin et al., 2017) was obtained from the Sina Weibo social platform, verified by the official Weibo rumor system, and then manually labeled by the Chinese Xinhua News Agency. This Weibo dataset of true information is collected from authoritative Chinese sources, and fake information is obtained through the official Weibo rumor suppression system. To assure the homogeneity of the entire dataset, we employ a similar strategy (Wang et al., 2018) to data preprocessing in this study to eliminate duplicate and poor-quality images. The dataset is split into training, validation, and test sets in the ratios of 7:1:2 to guarantee that no duplicate events are present in any of them.

4.1.2. Twitter dataset

Twitter dataset used for multimodal rumor detection was provided by Boididou et al. (2015). It consists of a development, which is the content of each tweet has a brief text message and extra images or videos. There are around 6,000 rumors and 5,000 non-rumor tweets in the development set from 11 rumor-related events. The test set includes approximately 2,000 tweets of both sorts. We excluded tweets with no text or images from this dataset because we only consider text and visual information (Zhang T. et al., 2020), Additionally, we use Google Translate to convert non-English information into English. The statistics of the two datasets are listed in Table 1.

4.2. Experimental setting

Due to the varied languages utilized in the two datasets, the model BERT is pre-trained on each language for the extracted text content. In the meantime, the settings of BERT and ResNet-50 are each frozen to prevent overfitting. The model is trained with 32-epoch batches and a learning rate of 0.001 across 100 iterations. To optimize parameters, the Adam optimizer is utilized.

4.3. Baseline models

Three groups of baseline models, single-modal, multi-modal, and MBPAM variants were selected to validate the effectiveness of MBPAM.

1. Single-modal models

- Text-GRU: Extract text features then input into the fully connected layer using the Bi-GRU model for classification.
- Image-VGG: Extract image features and feed them into the fully connected layer using the pre-trained VGG-19 network for classification.

2. Multi-modal models

- att-RNN (Jin et al., 2017): Using unidirectional LSTM and VGG-19, extract image features, context features and textual features. Then, employ RNN with attention to detecting fake news.
- EANN (Wang et al., 2018): Text-CNN model and VGG-19 model are applied to extract text and image features, which are then concatenated for utilization in the event discriminator and fake news detector.
- MVAE (Khattar et al., 2019): Using a bi-directional LSTM and VGG-19 to extract text and image features, which are then concatenated and fed to a self-encoder for reconstruction before being fed to a fake news detector.
- BDANN (Zhang T. et al., 2020): Using BERT model to get text features, while visual features are extracted using a VGG-19 model. Multimodal features of different events are mapped to the same space by using domain classifiers, removing dependency on specific events.
- Spotfake (Singhal et al., 2019): Employing BERT and VGG-19, respectively, to obtain text and image features, and then concatenate them into the fake news detector for detection.

3. Variants of MBPAM

- MBPAM-m: Use two branches to extract features of text and image separately, and then concatenate features into the event classifier and fake news detector.

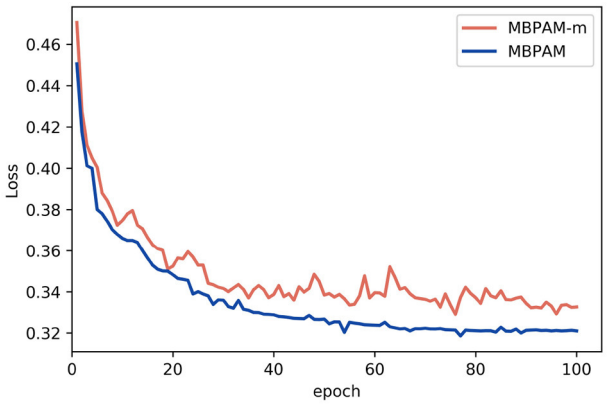

FIGURE 6
The comparison of MBPAM model and MBPAM-m's loss variation curve.

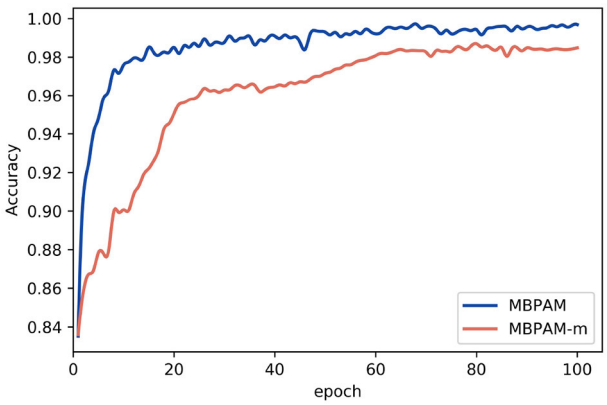

FIGURE 7
The comparison of MBPAM model and MBPAM-m's accuracy variation curve.

4.4. Results and analysis

4.4.1. Analysis of fake news detection

To analyze the impact of the fusion method in the MBPAM model on the training of the model, the loss value variation curves and accuracy variation curves of the MBPAM model and MBPAM-m model during training on the Weibo dataset are plotted.

As shown in Figure 6, from the perspective of stability, the loss curves of MBPAM with the fusion strategy are less oscillating and more stable. The loss curve of MBPAM with fusion strategy is less oscillating and more stable, while the MBPAM-m model is more oscillating and the curve is less stable. In terms of shrinkage, the fusion mechanism of MBPAM model speeds up the stabilization of the model.

From Figure 7, it can be analyzed that MBPAM, the fake news detection model with fusion strategy, has a smaller accuracy oscillation amplitude and more stable trend, while MBPAM-m also has a smaller accuracy oscillation amplitude but the trend is not stable enough. Also in terms of shrinkage speed, the MBPAM model with the fusion mechanism stabilizes earlier, which allows the model to complete the training task faster. This allows the model to complete the training task faster.

We carry out extensive experiments on two datasets to verify the performance of the MBPAM model. The evaluation standards include accuracy, precision, recall, and F1 score. Table 2 displays experimental results.

The experimental result demonstrates the MBPAM model outperforms the benchmark models in terms of accuracy. This indicates that the multimodal approach proposed in this study can significantly improve the detection of fake news. Especially with regard to accuracy, our methods are superior to those of other models. This indicates that our model is able to fully pay attention to the information between different modalities, fuse features and enhance features, which ultimately improves the accuracy of news classification. More concrete results are listed later.

First, it can be observed that the performance of the multimodal detection method on both datasets is better than that of the single-modal method, which proves that the image contain valid features that do not exist in the text, and the combination of two modalities achieves an effective information complementarity, which obtains richer depth features and improves the detection effect. Second, from the perspective of feature extraction, for the Weibo dataset, the accuracy of MBPAM-m is 9.8% higher than EANN and 8.7%

TABLE 2 Performance of MBPAM vs. other methods on Weibo and Twitter datasets.

Dataset	Method	Accuracy	Fake news			Real news		
			Precision	Recall	F1	Precision	Recall	F1
Weibo	Text-GRU	0.643	0.662	0.578	0.617	0.662	0.578	0.617
	Image-VGG	0.633	0.630	0.500	0.550	0.630	0.750	0.690
	att-RNN	0.772	0.797	0.713	0.692	0.684	0.840	0.754
	EANN	0.816	0.820	0.820	0.820	0.810	0.810	0.810
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	BDANN	0.842	0.830	0.870	0.850	0.850	0.820	0.830
	Spotfake	0.892	0.902	**0.964**	**0.932**	0.847	0.656	0.739
	MBPAM-m	0.896	0.923	0.874	0.898	0.868	0.920	0.893
	MBPAM	**0.904**	**0.943**	0.871	0.905	**0.868**	**0.941**	**0.903**
Twitter	Text-GRU	0.526	0.586	0.553	0.569	0.469	0.526	0.496
	Image-VGG	0.596	0.695	0.518	0.593	0.550	0.700	0.599
	att-RNN	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN	0.719	0.642	0.474	0.545	0.771	0.870	0.817
	MVAE	0.745	0.801	0.719	0.758	0.686	0.777	0.730
	BDANN	0.830	0.810	0.630	0.710	0.830	**0.930**	0.880
	Spotfake	0.777	0.751	**0.900**	**0.820**	0.832	0.606	0.701
	MBPAM-m	0.847	0.821	0.684	0.746	0.856	0.927	0.890
	MBPAM	**0.868**	**0.831**	0.754	0.791	**0.884**	0.925	**0.903**

The bold values indicate MBPAM-m and MBPAM. The bold part of the number represents the model that performs best on the two datasets under a certain evaluation criterion, that is, the model with the highest score.

higher than that of MVAE. The same trends on the Twitter dataset. Although EANN adds an event classifier, and MVAE is combined with a multimodal variational auto-encoder, none of them take into account the hidden layer text vector with contextual semantics and deep image vector information. Results have proved that the hidden layer information of the two branch extraction of text and images enriches the information of features, thus getting an improved performance.

Finally, from the point of view of feature fusion, the results of MBPAM are improved by 2.4% compared with MBPAM-m on the Twitter dataset, and about 1% higher than that of the Weibo dataset, which proved the effectiveness of the feature fusion module and feature enhancement module. At the same time, MBPAM on the Twitter dataset gains 11.7% and 4.5% improvements in accuracy over Spotfake and BDANN, respectively. On the Weibo dataset, the output benefit over Spotfake and BDANN is 1.3% and 7.3%, respectively. Although both Spotfake and BDANN use pre-trained models to extract vectors with contextual meaning and deep image vectors, which has improved the accuracy somewhat compared with the previous multimodal model, they do not pay attention to the information fusion among modalities, as they just simply concatenate them together for detection. Also, the results of MBPAM are greatly improved compared with att-RNN, MVAE, and EANN. So, both inter-modal feature information and intra-modal feature information are better represented in this study, which has shown that the multimodal bilinear pooling feature fusion and self-attention mechanism methods in our study are obviously better than the traditional vector concatenate method.

4.4.2. Analysis of ablation experiments

To evaluate the validity of the inter-modal feature fusion module and the intra-modal feature improvement module, ablation experiments are carried out on the Weibo dataset and the Twitter dataset.

- Remove the inter-modal feature fusion module: The text feature and image feature are gained, then concatenated and sent to the self-attention module, fake news detectors, and domain adversarial module. MBPAM-1 is the name of the model.
- Remove the intra-modal feature enhancement module: The text and image features are extracted and fed into multi-modal bilinear pooling feature fusion. The model is defined as MBPAM-2.
- Remove the domain adversarial module: Text and image features are extracted and fed into the multimodal feature fusion module for fusion and then directly into the fake news detector for detection. The model is defined as MBPAM-3.

Figure 8 illustrates the outcomes of the ablation experiment. When looking at the Weibo dataset, the results demonstrate that the accuracy of the model drops below that of the MBPAM model when the inter-modal feature fusion module is removed. It indicates that there is an information loss when text modality does not interact with image modality, which proves the importance of mining the relationship between different modalities for fake news detection and the effectiveness of multimodal bilinear pooling module fusion. However, the number of pictures in the Twitter

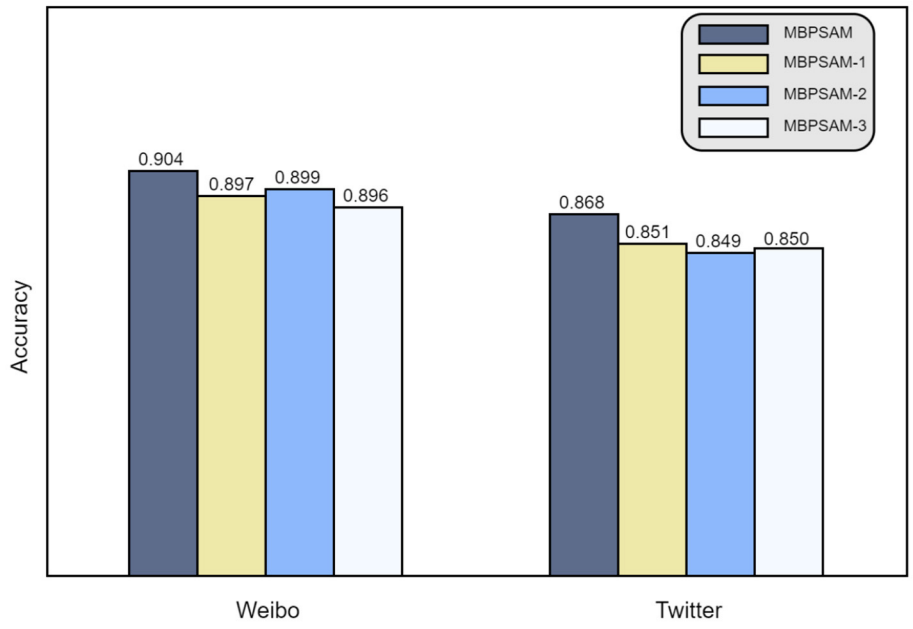

FIGURE 8
The ablation experiment on Weibo Dataset and Twitter Dataset. MBPAm-1 represents removing the inter-modal feature fusion module, MBPAm-2 represents removing the intra-modal feature enhancement module, and MBPAm-3 represents removing the domain adversarial module.

dataset is only about 510, and there is a phenomenon that one image corresponds to multiple news, resulting in insufficient image features in the Twitter dataset, while in the Weibo dataset, one text corresponds to one image. This difference may affect the information interaction of text and image, thereby affecting the fusion results. Therefore, when the inter-modal fusion module is used, the accuracy improvement in the Twitter dataset is not significant.

Second, when the intra-modal feature enhancement module is used in the model, the model's accuracy is improved on both datasets, indicating that the image features and text features are strengthened. If the interrelationship between the modalities is not sufficient, the use of intra-modal feature enhancement can effectively represent the feature information of each modality, as well as prevent the loss of features, thus ensuring the accuracy of detection. It also demonstrates that the intra-modal feature enhancement module is helpful for fake news detection. Therefore, the inter-modal feature fusion module and the intra-modal feature enhancement module really help to improve the detection effect of fake news. Finally, when the domain adversarial exists in the MBPAM model, the accuracy of the model is improved by about 1% on both datasets compared to MBPAM-3, proving that the existence of the domain adversarial mechanism is more helpful in finding generalized features and makes the model more generalizable.

5. Conclusion

In this study, we propose a multimodal fake news detection model based on self-attention mechanisms and multimodal bilinear pooling to handle the problem of merging text and image features for fake news detection. Two-branch networks and pre-training models are used in feature extraction to extract features and generate more useful data. The inter-modal information fusion module, which is based on multi-modal bilinear pooling, is used in feature fusion to merge the differences between text modality and image modality. The intra-modal information enhancement module, which is based on self-attention mechanism, is employed to give importance to important details within the modality. Through extensive experiments on two multi-modal datasets, the experimental results verify the effectiveness of the feature extraction module and fusion module, and the detection accuracy of our model is better than that of the benchmark model, since previous studies have focused only on the content of fake news and ignored the social subjects. However, fake news is generated by social subjects, so analyzing the features of social subjects is helpful for them to detect news. Also in the same post, sometimes several different images are attached, and different images convey information to users from multiple perspectives. In our future work, we will consider information on social subjects and think about how to combine textual information with multiple different images to enhance the accuracy of fake news detection.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

YG and JL contributed to the conception of the study. YG and HG performed the experiment and contributed significantly to analysis and manuscript preparation, performed the data analyses and wrote the manuscript, and helped perform the analysis with constructive discussions. All authors contributed to the article and approved the submitted version.

Funding

This research work was funded by the Beijing Social Science Foundation (21XCCC013).

Acknowledgments

The authors would like to express their gratitude for the support from the Beijing Social Science Foundation (21XCCC013). The authors would like to thank the reviewers for their constructive feedback and all the users who participated in this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., et al. (2015). "Vqa: visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision* (Santiago: IEEE), 2425–2433.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint* arXiv:1409.0473. doi: 10.48550/arXiv.1409.0473

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* 5, 157–166. doi: 10.1109/72.279181

Boididou, C., Andreadou, K., Papadopoulos, S., Dang-Nguyen, D.-T., Boato, G., Riegler, M., et al. (2015). Verifying multimedia use at mediaeval 2015. *MediaEval* 3, 7.

Castillo, C., Mendoza, M., and Poblete, B. (2011). "Information credibility on Twitter," in *Proceedings of the 20th International Conference on World Wide Web*, 675–684.

Charikar, M., Chen, K., and Farach-Colton, M. (2002). "Finding frequent items in data streams," in *International Colloquium on Automata, Languages, and Programming* (Springer), 693–703.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., and Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint* arXiv:1606.01847. doi: 10.18653/v1/D16-1044

Ganin, Y., and Lempitsky, V. (2015). "Unsupervised domain adaptation by backpropagation," in *International Conference on Machine Learning* (Lille: PMLR), 1180–1189.

Gao, J., Li, P., Chen, Z., and Zhang, J. (2020). A survey on deep learning for multimodal data fusion. *Neural Comput.* 32, 829–864. doi: 10.1162/neco_a_01273

Graves, A., Wayne, G., and Danihelka, I. (2014). Neural turing machines. *arXiv preprint* arXiv:1410.5401. doi: 10.48550/arXiv.1410.5401

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 770–778.

Horne, L., Matti, M., Pourjafar, P., and Wang, Z. (2020). "Grubert: A GRU-based method to fuse BERT hidden layers for Twitter sentiment analysis," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop* (Suzhou: Association for Computational Linguistics), 130–138.

Jin, Z., Cao, J., Guo, H., Zhang, Y., and Luo, J. (2017). "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proceedings of the 25th ACM International Conference on Multimedia* (New York, NY: ACM), 795–816.

Jin, Z., Cao, J., Zhang, Y., Zhou, J., and Tian, Q. (2016). Novel visual and statistical image features for microblogs news verification. *IEEE Trans. Multimedia* 19, 598–608. doi: 10.1109/TMM.2016.2617078

Khattar, D., Goud, J. S., Gupta, M., and Varma, V. (2019). "Mvae: multimodal variational autoencoder for fake news detection," in *The World Wide Web Conference* (New York, NY: ACM), 2915–2921.

Lahat, D., Adali, T., and Jutten, C. (2015). Multimodal data fusion: an overview of methods, challenges, and prospects. *Proc. IEEE* 103, 1449–1477. doi: 10.1109/JPROC.2015.2460697

Lin, T.-Y., RoyChowdhury, A., and Maji, S. (2015). "Bilinear cnn models for fine-grained visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision* (Santiago: IEEE), 1449–1457.

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.*

Ma, J., Gao, W., and Wong, K.-F. (2019). "Detect rumors on twitter by promoting information campaigns with generative adversarial learning," in *The World Wide Web Conference* (New York, NY: ACM), 3049–3055.

Meel, P., and Vishwakarma, D. K. (2020). Fake news, rumor, information pollution in social media and web: a contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Syst. Appl.* 153:112986. doi: 10.1016/j.eswa.2019.112986

Pham, N., and Pagh, R. (2013). "Fast and scalable polynomial kernels via explicit feature maps," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 239–247.

Qazvinian, V., Rosengren, E., Radev, D., and Mei, Q. (2011). "Rumor has it: Identifying misinformation in microblogs," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (Edinburgh: Association for Computational Linguistics), 1589–1599.

Qi, P., Cao, J., Yang, T., Guo, J., and Li, J. (2019). "Exploiting multi-domain visual information for fake news detection," in *2019 IEEE International Conference on Data Mining (ICDM)* (Beijing: IEEE), 518–527.

Salvi, C., Iannello, P., McClay, M., Rago, S., Dunsmoor, J. E., Antonietti, A., et al. (2021). Going viral: how fear, socio-cognitive polarization and problem-solving influence fake news detection and proliferation during COVID-19 pandemic. *Front. Commun.* 5, 562588. doi: 10.3389/fcomm.2020.562588

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: a data mining perspective. *ACM SIGKDD Explorat. Newsletter* 19, 22–36. doi: 10.1145/3137597.3137600

Singhal, S., Shah, R. R., Chakraborty, T., Kumaraguru, P., and Satoh, S. (2019). "Spotfake: a multi-modal framework for fake news detection," in *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)* (IEEE), 39–47.

Tenenbaum, J. B., and Freeman, W. T. (2000). Separating style and content with bilinear models. *Neural Comput.* 12, 1247–1283. doi: 10.1162/089976600300015349

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems, Vol. 30*. New York, NY: Curran Associates, Inc.

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* (New York, NY: ACM), 849–857.

Wu, K., Yang, S., and Zhu, K. Q. (2015). "False rumors detection on sina weibo by propagation structures," in *2015 IEEE 31st International Conference on Data Engineering* (Seoul: IEEE), 651–662.

Zhang, C., Yang, Z., He, X., and Deng, L. (2020). Multimodal intelligence: representation learning, information fusion, and applications. *IEEE J. Sel. Top. Signal Process.* 14, 478–493. doi: 10.1109/JSTSP.2020.2987728

Zhang, H., Fang, Q., Qian, S., and Xu, C. (2019). "Multi-modal knowledge-aware event memory network for social media rumor detection," in *Proceedings of the 27th ACM International Conference on Multimedia* (New York, NY: ACM), 1942–1951.

Zhang, T., Wang, D., Chen, H., Zeng, Z., Guo, W., Miao, C., et al. (2020). "BDANN: Bert-based domain adaptation neural network for multi-modal fake news detection," in *2020 International Joint conference on Neural Networks (IJCNN)* (Glasgow, UK: IEEE), 1–8.