Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.046202

ARTICLE

Multimodal Social Media Fake News Detection Based on Similarity Inference and Adversarial Networks

Fangfang Shan[1,2,*], Huifang Sun[1,2] and Mengyi Wang[1,2]

[1]College of Computer, Zhongyuan University of Technology, Zhengzhou, 450007, China

[2]Henan Key Laboratory of Cyberspace Situation Awareness, Zhengzhou 450001, China

*Corresponding Author: Fangfang Shan. Email: shanxiaofang1984@126.com

Received: 22 September 2023 Accepted: 23 February 2024 Published: 25 April 2024

ABSTRACT

As social networks become increasingly complex, contemporary fake news often includes textual descriptions of events accompanied by corresponding images or videos. Fake news in multiple modalities is more likely to create a misleading perception among users. While early research primarily focused on text-based features for fake news detection mechanisms, there has been relatively limited exploration of learning shared representations in multimodal (text and visual) contexts. To address these limitations, this paper introduces a multimodal model for detecting fake news, which relies on similarity reasoning and adversarial networks. The model employs Bidirectional Encoder Representation from Transformers (BERT) and Text Convolutional Neural Network (Text-CNN) for extracting textual features while utilizing the pre-trained Visual Geometry Group 19-layer (VGG-19) to extract visual features. Subsequently, the model establishes similarity representations between the textual features extracted by Text-CNN and visual features through similarity learning and reasoning. Finally, these features are fused to enhance the accuracy of fake news detection, and adversarial networks have been employed to investigate the relationship between fake news and events. This paper validates the proposed model using publicly available multimodal datasets from Weibo and Twitter. Experimental results demonstrate that our proposed approach achieves superior performance on Twitter, with an accuracy of 86%, surpassing traditional unimodal modal models and existing multimodal models. In contrast, the overall better performance of our model on the Weibo dataset surpasses the benchmark models across multiple metrics. The application of similarity reasoning and adversarial networks in multimodal fake news detection significantly enhances detection effectiveness in this paper. However, current research is limited to the fusion of only text and image modalities. Future research directions should aim to further integrate features from additional modalities to comprehensively represent the multifaceted information of fake news.

KEYWORDS

Fake news detection; attention mechanism; image-text similarity; multimodal feature fusion

1 Introduction

With the rapid development of mobile Internet technology, the primary platform for accessing news has shifted from traditional paper-based media, such as newspapers, to social media platforms

represented by Twitter and Weibo [1]. The real-time and convenient nature of social media enables people to quickly access and disseminate information. However, in the absence of effective supervision, the openness and low entry barriers of social media also facilitate the simultaneous spread of fake news [2]. Fake news is characterized by its low cost, rapid dissemination, and easy accessibility, which can lead to serious social issues, spark public opinion storms, and even manipulate public events, thereby undermining the credibility of social media. The rapid expansion of social media has become a breeding ground for the dissemination of fake news, where various fake news spread widely. Fake news not only has the potential to mislead the public but can also cause harm to individuals, organizations, and society. For instance, in 2019, Reuters reported that Hong Kong's Chief Executive, Carrie Lam, submitted a report to Beijing recommending the consideration of the 'five major demands' of Hong Kong's protesters. However, the report was allegedly rejected by Beijing. This fabricated news aimed to incite radical demonstrators, escalating the turmoil in Hong Kong, undermining the government's relationship with the people, and disrupting social harmony in China. Additionally, in 2020, amid the global COVID-19 pandemic, it became one of the primary sources of fake news [3].

To foster a harmonious online environment and mitigate the negative impact of fake news, there is an urgent need for reliable methods and technologies to address the issue of false information dissemination. Consequently, the detection of multimodal fake news on social media has emerged as a prominent research focus in recent years. Generally, there is no clear definition of fake news. The Merriam-Webster Online Dictionary defines fake news as "a news report that is intentionally false or misleading". Shu et al. [4] defined fake news as "false information that is intentionally misleading to readers and can be verified". Ajao et al. [5] defined fake news as "anything that is circulated, shared, or spread that cannot be authenticated". However, in academic research, fake news is usually defined as unverified or unconfirmed news. In this study, fake news is defined as intentionally misleading information that has been confirmed as false [6,7].

At present, the detection of fake news is primarily divided into two main directions: Unimodal modal detection and multimodal detection. Unimodal modal methods rely solely on text or image features for fake news detection. However, a news story embodies falseness in both text and image aspects, limiting the effectiveness of this approach in capturing the diverse features of fake news. In contrast, multimodal detection integrates features from various modalities, such as text and images, allowing for a more comprehensive understanding of fake news content.

Current multimodal fake news detection methods typically connect textual and visual features to obtain a unified multimodal feature representation. Nevertheless, these methods have yet to fully explore the similarity relationships between multimodal information, which is crucial for accurate fake news detection. Some fake news stories, aiming to garner clicks and widespread dissemination, often employ provocative image information that deviates significantly from the actual news text. For instance, Fig. 1 illustrates a fake news story about the U.S. government purchasing 30,000 guillotines, where the accompanying image features a historical painting depicting the beheading of Queen Marie Antoinette, creating a discordance with the textual content.

Addressing the semantic bias between textual and visual content, this paper introduces a multimodal social media fake news detection approach grounded in similarity reasoning and adversarial networks. Specifically tailored to bridge the research gap in understanding the similarity relationships within multimodal information, our method aims to comprehensively and accurately unveil the characteristics of fake news. By delving deeply into the correlations between text and images, our approach provides a nuanced and precise perspective for fake news detection. The method comprises five modules: (1) multimodal feature extractor; (2) similarity representation learning and reasoning; (3)

multimodal feature fusion; (4) fake news detector; (5) event classifier. Designed to discern the falseness of news articles in terms of text, image, or "mismatch". Our main contributions are summarized as follows:

(1) We conduct a comprehensive consideration of both local and global features in text and images. To extract text features, we integrate BERT and Text-CNN, introducing an attention mechanism post-Text-CNN to capture global text features. Simultaneously, leveraging the pre-trained VGG-19 on ImageNet, we extract local features from images. Following VGG-19, an attention mechanism is incorporated to capture global image features.

(2) We employ similarity representation learning and inference to infer the similarity between images and text, thereby recognizing more intricate matching patterns.

(3) By integrating event-based adversarial networks with multimodal networks, we not only capture features specific to particular events but also learn the associations between modal features and events.

(4) Extensive experiments on publicly available multimodal datasets, the research results demonstrate the outstanding performance of the proposed model in fake news detection tasks, particularly on the Twitter dataset. In comparison to traditional detection models, this model consistently achieves superior results across multiple metrics, effectively enhancing the accuracy and performance of fake news identification.

Figure 1: Example of inconsistent graphic content of fake news

2 Related Work

In this section, we present the research related to the proposed model for detecting fake news. Fake news detection has been a widely researched area due to its significance in maintaining the accuracy and reliability of public information. Existing fake news detection methods can be classified into two categories: Unimodal-based and multimodal-based fake news detection.

2.1 Unimodal-Based Fake News Detection

In the domain of unimodal-based fake news detection, the prevailing strategy revolves around leveraging textual information to ascertain the authenticity of news articles. This approach involves evaluating text content, syntactic structures, themes, and other factors to establish the veracity of the news. In the study conducted by Liu [8], the TF-IDF algorithm is applied to extract text features, and these features are utilized as inputs for a Support Vector Machine (SVM) to distinguish the authenticity of news. In contrast to certain intricate deep learning models, the TF-IDF algorithm

exhibits superior computational efficiency, particularly when handling extensive textual data. Nevertheless, it falls short in capturing the contextual relationships between words, a crucial aspect for precise determinations of information authenticity in fake news detection. Amid the advancements in deep learning technology, models based on neural networks now can acquire more profound and abstract features, enabling end-to-end learning. In the context of fake news detection, Ma et al. [9] employed Recurrent Neural Networks (RNNs) to input all news texts associated with a specific event. The final hidden state of the RNN was subsequently leveraged to discern fake news at the event level. Although RNNs demonstrate proficiency in capturing contextual information within text sequences, they confront challenges in capturing long-distance dependencies when handling extensive sequences. Addressing the early detection challenge in fake news, Yu et al. [10] presented a Convolutional Neural Network (CNN)-based approach. This method initially groups news about the same event, transforming the textual content of each newsgroup into a document vector. Subsequently, CNN extracts text features from multiple document vectors for fake news detection. In comparison to models like RNN, which necessitate the consideration of sequence information, CNN's efficiency in text processing is notable for not relying on sequence information. However, CNN is typically employed to capture local features, with limited capacity for processing global information. Ma et al. [11] further improved the model's performance by introducing generative adversarial networks to enhance the learning of textual representations in fake news detection. Generative Adversarial Networks enable enhanced and nuanced learning of text representations. The interplay between the generator and discriminator allows the model to acquire text representations that are both distinctive and abstract. It is noteworthy, however, that the application of GANs often necessitates considerable computing resources, such as high-performance GPUs, and entails prolonged training time. While the success of text feature-based fake news detection is evident in certain aspects, the consideration of textual features alone lacks comprehensiveness, as fake news may be accompanied by misleading images or charts.

With the continuous advancement of image processing technology, the decreasing difficulty in forging false images poses greater challenges for the general public in discerning the authenticity of news and, consequently, presents a more significant challenge for fake news detection [12]. Therefore, scholars have increasingly focused on the detection of fake images. Mahmood et al. [13] proposed a method that combines the smooth wavelet transform and Discrete Cosine Transform (DCT) to detect and locate copy-move operations in images. This method comprehensively captures image features in the frequency domain, aiding in the more precise detection of copy-move operations. However, it encounters challenges in effectively addressing intricate textures or semantic information. Farooq et al. [14], combining Local Binary Pattern (LBP) features and texture features, introduced a method using a universal algorithm based on LBP to detect passive image forgery. While this method adeptly captures local texture information within images, its comprehension of global structure and context is constrained. In situations involving intricate forgery techniques, it lacks the requisite discriminative capacity. Peng et al. [15] identified forged images by examining resampling traces in the images. In contrast to approaches that require a reference image for comparison, this method does not necessitate obtaining a reference image beforehand, making it practical for real-world applications. However, it is typically employed for the overall assessment of whether an image is manipulated and does not offer detailed localization of forged regions. Zeng et al. [16] employed a hybrid deep-learning model to detect steganographic operations in JPEG images. The model utilizes techniques such as quantization and truncation to enhance its robustness and generalization capabilities. Quan et al. [17] developed a convolutional neural network-based universal model capable of classifying images into natural and computer-generated categories, making it suitable for various fake image detection scenarios.

Despite the promising results that unimodal modal methods can offer to some extent, data in social networks often involve multimodal information such as text and images. Unimodal modal methods fall short of adequately capturing and processing this diversity and complexity. As a result, researchers are beginning to explore the combination of text and images to address the limitations of unimodal methods.

2.2 Multimodal-Based Fake News Detection

Currently, deep neural networks (DNNs) excel in nonlinear representation [18], making them a prominent choice for many multimodal representation learning methods aimed at enhancing the capability of fake news detection. Jin et al. [19] proposed a deep learning-based approach capable of learning multimodal content and social information from news posts. They introduced an attention mechanism to fuse this information and obtain multimodal features, enhancing the model's focus on crucial information and improving the weight allocation for different modal data, allowing for more effective utilization of multimodal information. In EANN, Wang et al. [20] employed an adversarial network with a multimodal feature extractor to learn invariant features of events, acquiring multimodal features for each news article to facilitate fake news detection. Learning invariant features of events through adversarial networks enhances the model's generalization, yielding favorable results across diverse events. In MVAE, Khattar et al. [21] utilized a multimodal variational autoencoder for fake news identification, inputting various modal features of posts into a bimodal variational autoencoder to obtain multimodal feature representations. The introduction of variational autoencoders aids in learning latent representations of data, thereby enhancing the model's expressive and generalization capabilities. Cui et al. [22] introduced an end-to-end deep embedding framework (SAME) for fake news detection. In this model, the emotions of post publishers serve as the basis for discerning fake news. By embedding emotional features with other characteristics through deep learning, the model distinguishes between real and fake news. Leveraging post publishers' emotions as a basis for judging fake news provides additional information, contributing to a more comprehensive understanding of the authenticity of news. SpotFake [23] utilizes a pre-trained BERT [24] model for text feature extraction from news posts and employs a pre-trained VGG-19 model on ImageNet [25] for image feature extraction. The use of pre-trained BERT models enables the learning of rich text representations. SpotFake+ [26], an enhanced version of SpotFake, utilizes an improved BERT variant, XLNet [27] for text feature extraction.

Despite the current technological advancements propelling the development of multimodal fake news detection, there remains limited exploration and utilization of relationships between different modalities. This paper aims to address this gap by introducing similarity representation learning and inference, filling the research void in understanding the relationships between news text and visual information. By exploring multimodal information and similarity relationships, this study seeks to comprehensively understand and learn the representations of news articles. Additionally, the introduction of adversarial networks for learning invariant features of events aims to advance the frontier of research in multimodal fake news detection.

3 Method

3.1 Model Overview

This paper introduces a multimodal social media fake news detection model based on similarity reasoning and adversarial networks. The model comprises a multimodal feature extractor, similarity representation learning and inference, multimodal feature fusion, a fake news detector, and an event

classifier. Initially, the model independently preprocesses text and images, subsequently extracting feature representations. Textual features are extracted using the BERT and Text-CNN models. Subsequently, an attention mechanism is introduced post-Text-CNN to capture the global features of the text. For image feature representation, the pre-trained VGG-19 model is employed to acquire local image features, followed by applying a Self-Attention mechanism to these local features to derive global image feature representations. The model learns local similarity representations for text and image through the Text-CNN model for text local features and VGG-19 for image local features. Simultaneously, it acquires global similarity representations for text and image through the global features extracted from text and images, respectively. All the local similarity representations and global similarity representations serve as nodes within a graph, and we calculate the edges connecting these nodes. The graph undergoes similarity reasoning, which involves updating the nodes and edges iteratively over the N steps of reasoning. The output of the global nodes from the final step is used as the inferred similarity representation. This representation is then passed through a fully connected layer to generate the ultimate similarity score. Concatenate the textual features extracted by the BERT model with the text local features extracted by the Text-CNN model to form the textual feature representation. Subsequently, perform feature fusion by concatenating this textual feature representation with the local image features extracted by VGG-19 and the similarity representation resulting from text-image inference. The event classifier is a neural network model with a structure consisting of two fully connected layers, each equipped with corresponding activation functions. It employs clustering algorithms to accurately categorize newly emerged news information into specific event classes. The computation loss of an event is indicative of the similarity of the event distributions, with a larger loss denoting a greater similarity. The fake news detector utilizes the Softmax function as the activation function for the output layer. This function transforms the output of the fully connected layer into activation values representing the probability of fake news. Through this mechanism, the model can classify input features, discerning whether the news is genuine or fake. The framework of the multimodal social media fake news detection model based on similarity reasoning and adversarial networks (EANBS) is illustrated in Fig. 2.

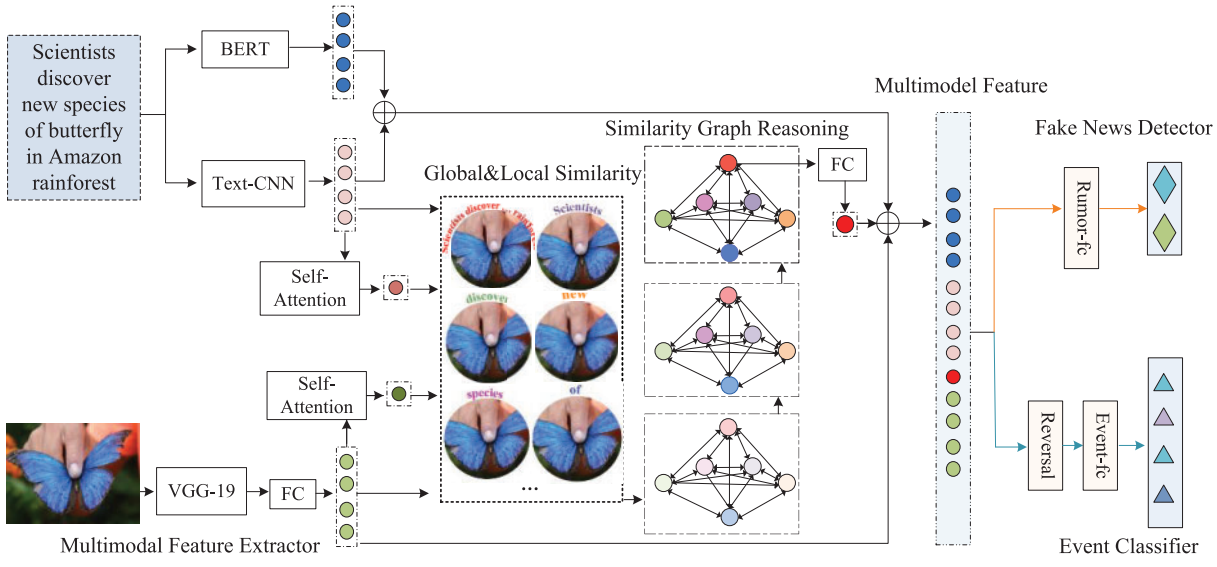


Figure 2: EANBS model structural framework

3.2 Multimodal Feature Extractor

3.2.1 Textual Feature Extractor

This paper utilizes two critical text feature extractors: The Text-CNN and the BERT model. The Text-CNN allows for a focused analysis of localized perspectives and fine-grained features within the text, while the BERT model excels at extracting deep-seated semantic characteristics from the text. The synergy between these two approaches enables a more efficient extraction of textual semantic features.

In Text-CNN, a convolutional layer is utilized to extract features at a local level. By applying convolutional operations, the model can capture nuanced characteristics within the text. This, in turn, facilitates a more comprehensive understanding of localized information present in the text, ultimately refining the representation of text features. The standard configuration of the Text-CNN model consists of an embedding layer, a convolutional layer, a pooling layer, and a fully connected layer. The arrangement of the Text-CNN model is visually depicted in Fig. 3.

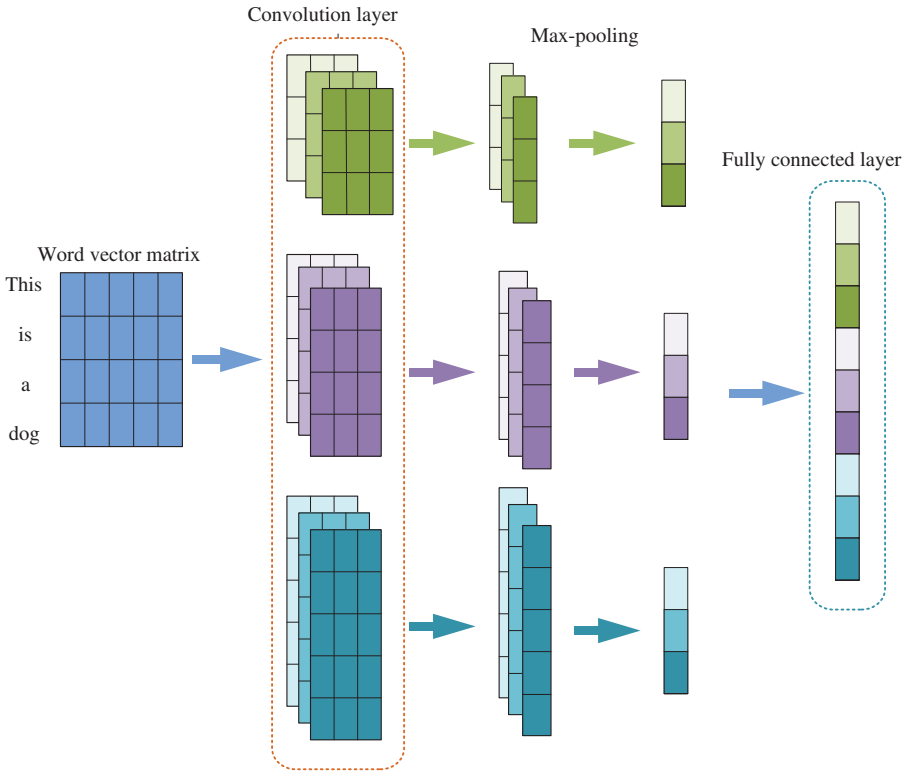


Figure 3: The structure of Text-CNN model

Local textual feature: (a) The Embedding layer: Words or phrases can be depicted as continuous, low-dimensional vectors. In the embedding layer, a sentence X consisting of m words can be represented as shown in Eq. (1).

$$X_{1:M} = X_1 \oplus X_2 \oplus \cdots \oplus X_m \tag{1}$$

where X_i represents the i-th word of the current text and $\oplus$ represents the vector splicing operation.

(b) The convolutional layer: After converting text into word vectors using Word2Vec, we employ CNN's convolutional operation to capture local features within the text. This is achieved by configuring various convolution kernel sizes s to process textual fragments of different lengths. The convolutional operation can be represented as shown in Eq. (2).

$$h_i = f\left(W \cdot X_{i:i+s-1}\right) \tag{2}$$

where W represents the weight matrix of the output convolution kernel, f is the activation function, and the vector composed of h_i is the feature vector extracted from the convolution layer, i.e., $h = \{h_1, h_2, \cdots, h_n\}$, which is taken as the input to the pooling layer.

(c) The pooling layer: Following the convolutional layer, Text-CNN typically employs a max-pooling operation to reduce the dimensionality of the output generated by the convolutional operation. This helps in extracting essential features from the text effectively, as illustrated in Eq. (3).

$$\overline{h} = max\left(h\right) \tag{3}$$

(d) The fully connected layer: Finally, the vector representation obtained from the preceding pooling layer is combined through the fully connected layer to yield the local text representation H.

Global textual feature: Incorporating the attention mechanism can help the neural network model focus more on essential words. This allows the model to prioritize information from these keywords while ignoring less significant segments. As a result, the influence of unnecessary data is reduced, enhancing the model's ability to extract crucial information. This, in turn, boosts the model's efficiency and accuracy.

The Self-Attention mechanism, a specific case of the general attention mechanism, stands out for its ability to learn intrinsic textual correlations. When combined with CNN or RNN, it significantly enhances the model's learning capabilities and improves the interpretability of the neural network. The computational steps of the Self-Attention mechanism can be represented as Eqs. (4)–(7).

$$Q = W^Q H \tag{4}$$

$$K = W^K H \tag{5}$$

$$V = W^V H \tag{6}$$

$$\overline{H} = Attention\left(Q, K, V\right) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{7}$$

where H represents the local textual features, Q is the query vector, K is the "key" vector, V is the value vector, and W^Q、W^K、W^V is the corresponding weight matrices. $\overline{H}$ represents the global textual features.

BERT, a pre-trained language representation model, is constructed with stacked bidirectional Transformer [28] encoder structures. In contrast to traditional unidirectional models, BERT comprehensively captures contextual nuances in text, proving particularly advantageous for intricate textual scenarios encountered in tasks such as fake news detection. Leveraging extensive unsupervised pre-training, BERT acquires rich language representations encompassing a broad spectrum of semantic knowledge. The incorporation of these pre-trained weights into our task endows the model with the ability to benefit from a wealth of prior knowledge, thereby enhancing its capacity to articulate text features. Through a self-attention mechanism, BERT adeptly learns global contextual information, facilitating a deeper understanding of semantic relationships across the entire text. This results in more comprehensive word vector representations. Consequently, we integrate Text-CNN and BERT as the two core modules for text feature extraction. First, tokenize the input text, converting the tokens into

word embeddings and positional embeddings. Subsequently, input these embeddings into the BERT model. The feature extraction process by BERT is outlined in Eqs. (8), (9).

$$B = BERT\,(X) \tag{8}$$

$$B = B \cdot W_1 \tag{9}$$

where X represents the input text, and W_1 is the weight matrix of the fully connected layers in the corresponding pre-trained model.

To synergistically leverage the advantages of both Text-CNN and BERT models, this study adopts a concatenation approach, as illustrated in Eq. (10).

$$T = H \oplus B \tag{10}$$

Specifically, the text features extracted by BERT are concatenated with the local features extracted by Text-CNN to form the final representation of text features. This fusion strategy aims to integrate global and local information, enabling a more comprehensive and multidimensional expression of features in news text.

In the context of fake news detection tasks, the significance of this fusion strategy lies in its ability to enhance the model's comprehension of complex, multilayered information. By integrating both global and local features, the model becomes adept at distinguishing between authentic and fake news, as it can more comprehensively capture the semantic and structural information embedded in news text.

3.2.2 Visual Feature Extractor

In general, the brain learns and comprehends visual information much faster than textual information. Based on this insight, we have considered the visual features of news. Integrating the extracted image features with textual features enhances the feature representation, leading to a more comprehensive understanding and assessment of fake news. VGG-19 stands as a classical convolutional neural network architecture, exhibiting exceptional performance in image classification tasks. The hierarchical structure of VGG-19, with its relative depth, facilitates the extraction of abstract features from images, demonstrating excellent adaptability to the intricate patterns and structures potentially present in images associated with fake news. Pre-training on large-scale image datasets endows VGG-19 with a rich feature representation. Leveraging these pre-trained weights allows the model to learn universal visual features from diverse images, a significant advantage in the context of fake news image detection. Therefore, this paper employs the pre-trained VGG-19 model for extracting image features. To better preserve image information, resize the input image to 256 $\times$ 256 pixels. Perform a center crop to reduce it to 224 $\times$ 224 pixels. Preprocess the cropped image and feed it into the VGG-19 model. Add a fully connected layer at the end, adjusting the final image feature dimension to c, serving as the representation of the image's local regions $I = \{I_1, I_2, \cdots, I_n\}$, with $I_i \in R^c$. The above can be represented in Eqs. (11), (12).

$$I = Vgg\,(g) \tag{11}$$

$$I = I \cdot W_2 \tag{12}$$

where g represents the input image and W_2 is the weight matrix of the fully connected layer in the corresponding pre-trained model. Then the self-attention mechanism is utilized to derive the global

image region features by Eqs. (13)–(16).

$$Q = W^Q I \tag{13}$$

$$K = W^K I \tag{14}$$

$$V = W^V I \tag{15}$$

$$\bar{I} = Attention\,(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{16}$$

where I represents the local textual feature, Q is the query vector, K is the "key" vector, V is the value vector, and W^Q、W^K、W^V are the corresponding weight matrices. $\bar{I}$ represents the global visual features. The choice of Text-CNN, BERT, and VGG-19 as feature extractors is driven by their outstanding performance in their respective domains and their complementary characteristics. This selection aims to enhance the performance of fake news detection.

3.3 Similarity Representation Learning and Reasoning

In the realm of fake news detection, traditional methodologies often focus solely on extracting text and image features, disregarding the inherent parallels between the two. However, comprehending this intrinsic similarity is paramount for accurate fake news prediction. Thus, the introduction of similarity representation learning and inference emerges as a crucial avenue. This approach facilitates the absorption of shared semantic nuances between text and image, consequently elevating the accuracy and robustness of fake news detection.

3.3.1 Similarity Representation Learning

Traditional methods utilize the cosine or Euclidean distance to represent the similarity between two feature vectors, which can capture the relevance to a certain degree while lacking detailed correspondence. In this paper, we follow [29] to compute a similarity representation, which is a similarity vector instead of a similarity scalar, to capture more detailed associations between feature representations from different modalities. The similarity function representation can be represented by Eqs. (17), (18).

$$d\,(x, y) = W\,|x - y|^2 \tag{17}$$

$$sim\,(d) = \frac{d}{\parallel d \parallel_2} \tag{18}$$

where $x \in R^d$, $y \in R^d$, $|\cdot|^2$ is the element-wise square, $\parallel \cdot \parallel_2$ is the l_2 normalization, and $W \in R^{q \times d}$ is a learnable parameter matrix to obtain a q-dimensional similarity vector, $sim\,(d)$ is the similarity function.

Global Similarity Representation. The global similarity, $sim^g = \frac{d}{\parallel d \parallel_2}$ with $d = Wg\left|\bar{I} - \overline{H}\right|^2$, is derived from the global image feature $\bar{I}$ and sentence features $\overline{H}$ through the utilization of Eqs. (17),(18). The parameter matrix $Wg \in \mathrm{R}^{q \times d}$ designed for learnability, serves the purpose of capturing the global similarity representation.

Local Similarity Representation. To calculate the local similarity representations between local features found in visual and textual observations, apply textual-to-visual attention [30], calculate the

cosine similarity between the region feature I_i with to word feature H_j first by Eq. (19).

$$s_{ij} = \frac{I_i^T H_j}{I_i H_j}, i \in [1, m], j \in [1, n] \tag{19}$$

The cosine similarity matrix is then normalized and can be represented in Eq. (20).

$$\overline{s_{ij}} = [s_{ij}]_+ / \sqrt{\sum_{i=1}^{m} [s_{ij}]_+^2}, [x]+ \equiv max(x, 0) \tag{20}$$

Next, we calculate the similarity between the word feature H_j and the entire visual features. Compute attention weight for each region, then we generate visual features $a^I{}_j$ concerning the j-th word by Eqs. (21), (22).

$$\alpha_{ij} = \frac{exp(\lambda \overline{s_{ij}})}{\sum_{i=1}^{n} exp(\lambda \overline{s_{ij}})} \tag{21}$$

$$a^I{}_j = \sum_{i=1}^{n} \alpha_{ij} I_i \tag{22}$$

where α_{ij} represents the attention weight of each region.

The local similarity representation $sim^l{}_j = \frac{d}{\| d \|_2}$ between $a^I{}_j$ and H_j can be computed using Eqs. (17), (18), where $d = W_3 \left| a^I{}_j - H_j \right|^2$. The parameter matrix $W_3 \in R^{q \times d}$ designed for learnability, serves the purpose of capturing the local similarity representation.

3.3.2 Similarity Reasoning

To propagate similarity information among possible alignments at both local and global levels, a similarity graph is constructed with global similarity and local similarity as nodes. The relationship between two similarities serves as edges in the graph, i.e., $N = \{sim^l_1, \cdots, sim^l_L, sim^g\}$, and follow [31] to calculate the edge from node $sim_q \in N$ to $sim_p \in N$ as Eq. (23).

$$L(sim_p, sim_q; F_{in}, F_{out}) = \frac{exp((F_{in} sim_p)(F_{out} sim_q))}{\sum_q exp((F_{in} sim_p)(F_{out} sim_q))} \tag{23}$$

where $F_{in} \in R^{q \times q}$ and $F_{out} \in R^{q \times q}$ represent the linear transformations for incoming and outgoing nodes, respectively.

In this paper, the similarity is inferred through similarity propagation, linear transformation, and a non-linear activation function. Specifically, the similarity is first propagated as Eq. (24).

$$\widetilde{sim^n_p} = \sum_q \mathrm{L}(sim^n_p, sim^n_q; F^n_{in}, F^n_{out}) \cdot sim^n_q \tag{24}$$

where sim^0_p and sim^0_q from N at step $n = 0$, and F^n_{in}, F^n_{out} are learnable parameters. Next, nonlinear activation using the ReLU function is performed during reasoning as Eq. (25).

$$sim^{n+1}_p = ReLU(W_t \widetilde{sim^n_p}) \tag{25}$$

where $W_t \in R^{q \times q}$ is a learnable parameter.

We iterate reasoning for N steps, and the output of the global node at the last step of iterative reasoning is utilized as the similarity representation of the reasoning and inputted into the fully

connected layer to obtain the final similarity score as Eq. (26).

$$S_f = fc\left(sim_q^{n+1}\right) \tag{26}$$

3.4 Multimodal Feature Fusion

The representations of textual features, visual features, and the similarity representations between text and image are combined to create a multimodal feature representation R_F in Eq. (27).

$$R_F = T \oplus I \oplus S_f \tag{27}$$

where T is the concatenation of textual features extracted by BERT and local features extracted by Text-CNN, I represents the local visual features extracted by VGG-19, and S_f represents the similarity representation between textual features extracted by Text-CNN and visual features extracted by VGG-19. This concatenation method enables the retention of distinctive features from each modality, facilitating a better understanding of the interactions between text and images, thereby improving the performance in fake news detection tasks. The multimodal feature extractor is represented as $F\left(M;\theta_f\right)$, where M usually refers to a set of textual and visual posts, and θ_f represents the parameters to be learned.

3.5 Fake News Detector

This module is designed to implement a neural network for detecting fake news, which is built upon a multimodal feature extractor. Textual features, visual features, and their similar representations are combined to create a multimodal feature representation R_F. This representation is then used as input to the network and deploys a SoftMax fully connected layer for classification to predict whether the post is fake news. The fake news detector is denoted as $C\left(F;\theta_C\right)$. The probability that this post is a fake is shown in Eq. (28).

$$\hat{y} = C\left(F\left(m_i;\theta_f\right);\theta_C\right) \tag{28}$$

where θ_C represents all the parameters of this network, $\hat{y}$ represents the probability that the current post is fake news. The real news is labeled as 0, and the fake news is labeled as 1. Y is used to denote the true labels of news events and the detection loss is computed using sigmoid cross-entropy, as shown in Eq. (29).

$$L_C\left(\theta_f,\theta_c\right) = -E_{(m,y)\sim(M,Y)}\left[y\log\left(\hat{y}\right) + (1-y)\right]\log\left(1-\hat{y}\right) \tag{29}$$

where M usually refers to a set of textual and visual posts. We seek to minimize the loss in classifying fake news by searching for optimal parameters and this process can be represented in Eq. (30).

$$\widehat{\theta_f},\widehat{\theta_C} = argmin_{\theta_f,\theta_C} L_C \tag{30}$$

3.6 Event Classifier

The event classifier consists of two fully connected layers, and its role is to classify K events. It evaluates the performance and robustness of the feature extractor by comparing the differences between the feature representation and the original events. This approach eliminates the strict dependence on specific events in the collected dataset and enables better generalization to unseen events, which in turn guides the training of the feature extractor. The event classifier is denoted as $D\left(F;\theta_D\right)$, where θ_D represents its parameters. The loss of event classifier is defined by cross-entropy,

and the label set of events is denoted using Y_e. This process can be represented in Eq. (31).

$$L_D\left(\theta_f, \theta_D\right) = -E_{(m,y_e)\sim(M,Y_e)}\left[\sum_{k=1}^{K} y_e\left(D\left(F\left(m;\theta_f\right)\right);\theta_D\right)\right] \tag{31}$$

The minimization of the loss function is expressed as Eq. (32).

$$\widehat{\theta_D} = argmin_{\theta_D} L_D \tag{32}$$

By using the loss function L_D to measure similarities and differences between different events, a larger value of the loss function indicates that the distribution of different events is more similar. Therefore, the gradient descent method is used to find the optimal parameters θ_f to maximize the loss function L_D, allowing for better distinction between events and fake news, and discovering the association between them.

3.7 Model Integration

During training, minimizing loss L_C is crucial to enhance the model's ability to discern fake information and improve classification accuracy. To ensure that the model can effectively acquire shared event features, it is necessary to maximize the loss L_D of the event classifier. Simultaneously, the event classifier also strives to minimize loss L_C to extract event-specific information from multimodal feature representations. Consequently, the overall loss can be represented in Eq. (33).

$$L_Z\left(\theta_f, \theta_C, \theta_D\right) = L_C - \lambda L_D \tag{33}$$

where the coefficient $\lambda \in R$ is employed to strike a balance between the objective functions of the fake news detector and the event classifier. For the problem of the maximin game, this paper utilizes a Gradient Reversal Layer (GRL). During the forward pass, the Gradient Reversal Layer acts as an identity function, whereas during the backward pass, GRL multiplies the gradients $-\lambda$ and propagates them to the preceding layer. The parameter optimization process can be represented in Eq. (34).

$$\theta_f \leftarrow \theta_f - \mu\left(\frac{\partial L_C}{\partial \theta_f} - \lambda\frac{\partial L_D}{\partial \theta_f}\right), \theta_C \leftarrow \theta_C - \mu\frac{\partial L_C}{\partial \theta_C}, \theta_D \leftarrow \theta_D - \mu\frac{\partial L_D}{\partial \theta_D} \tag{34}$$

4 Experiment

4.1 Dataset

Weibo Dataset: The Weibo Dataset, as presented by Jin et al. [32], has been extensively employed in numerous studies focused on multimodal fake news. This dataset spans confirmed fake news from May 2012 to January 2016, officially verified by Weibo, and real news validated by China's authoritative news source, Xinhua News Agency. During the data preprocessing phase, a meticulous multi-step approach was applied to ensure dataset quality. Initially, duplicate images were removed to alleviate redundancy in the data. Subsequently, low-quality images were filtered out to ensure that all images in the dataset maintained high clarity and usability standards. Only data samples featuring both textual and image modalities were utilized to prevent distributional biases in unimodal and multimodal experiments, thereby enhancing the persuasiveness and credibility of the results. The dataset was partitioned into training, validation, and testing sets in a 7:1:2 ratio.

Twitter Dataset: The Twitter dataset utilized in this study is sourced from the MediaEval2015 dataset [33], encompassing both a training set and a test set. Each news item in the dataset consists of supplementary images/videos, text, and labels. In the data preprocessing stage, punctuation, numbers,

special characters, and short words were removed from the tweets. Given the emphasis of our work on textual and image information, tweets with videos were excluded. Examples of images and corresponding text in the dataset are illustrated in Figs. 4 and 5. The specific distribution of each dataset is presented in Table 1.

(a) New species of fish found in Arkansas

(b) There are more than 66 million people affected by the Nepal earthquake

Figure 4: Instances of fake news in the Twitter dataset

(a) Great picture Thierry Legault captures iss transit of the sun during eclipse

(b) View from our office right now

Figure 5: Instances of real news in the Twitter dataset

Table 1: Distribution of each dataset

Method	Weibo	Twitter
#Of fake news	4108	7875
#Of real news	3615	5220
#Of images	7723	410

4.2 Experimental Details

The specific model parameters for this experiment are detailed in Table 2. Throughout this paper, several experiments were conducted to optimize these parameter settings. By making repeated adjustments and conducting multiple experiments, we were able to identify the optimal parameter configurations. The Text-CNN filter window selects a fixed set of sizes [1,2,3,4], which serves to reduce

the hyperparameter search space of the model. This simplification alleviates the complexity and tedium associated with the hyperparameter tuning process. By capturing features at various levels, the model becomes better suited to adapt to text inputs of differing lengths and complexities. This adaptation contributes to the enhancement of classification accuracy and robustness across various scenarios.

Table 2: Model parameters

Parameters	Numerical value
Word embedding model dimensions	32
Dropout	0.5
Learning rate	0.001
Batch size	32
Epoch	50
Optimizer	Adam
Reasoning step N	3
Text-CNN filter window	[1,2,3,4]
Hidden layer size of the textual and visual feature extractor	32
Hidden layer size for detecting fake news	64
Hidden layer size of the first layer of event classifier	64
Hidden layer size of the second layer of event classifier	32

4.3 Evaluation Metrics

We used the traditional performance metrics namely accuracy, recall, precision, and F_1 score to evaluate the proposed model. Here is a brief explanation of these metrics.

$$Accuracy = \frac{TP + TN}{TP + TF + FP + FN} \tag{35}$$

$$Precision = \frac{TP}{TP + FP} \tag{36}$$

$$Recall = \frac{TP}{TP + FN} \tag{37}$$

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{38}$$

where True positive (TP): Fake news forecasted as fake; True negative (TN): Real news forecasted as real; False positive (FP): Real news forecasted as fake; False negative (FN): Fake news forecasted as real.

4.4 Baseline

To validate the effectiveness of the proposed model, this study selects two categories of baseline models: Unimodal models and multimodal models.

4.4.1 Unimodal Models

The unimodal model employs either textual or visual information alone to detect fake news. Thus, this paper proposes the following two sample baselines:

•**Textual.** The model exclusively relies on the textual content within the post for post-classification, utilizing a pre-trained 300-dimensional Word2Vec model from Sogou Labs to represent word vectors. Firstly, Text-CNN is employed for text feature extraction, transforming the textual information into a feature representation. Subsequently, the extracted text features undergo processing through a 32-dimensional fully connected layer to accomplish the task of detecting fake news.

•**Visual.** The model relies exclusively on the image information within the post for post-classification, starting by extracting the image feature, denoted as F_v, using a pre-trained VGG-19 model. Subsequently, the acquired image feature F_v is input into a 32-dimensional fully connected layer to make predictions regarding fake news.

4.4.2 Multimodal Models

Multimodal approaches utilize information from multiple modalities to classify fake news.

•**VQA [34].** Visual Question and Answer (VQA) is a system that provides answers to questions based on a given image. While the original VQA model was designed for a multiclassification task, the primary focus of this paper is on binary classification.

•**NeuralTalk [35].** NeuralTalk is a model designed to generate subtitles for a given image. It obtains potential representations of the text sequence by averaging the output of the RNN (Recurrent Neural Network) at each time step. These latent representations are then passed to a fully connected layer for the prediction of fake news. Both the LSTM and the fully connected layer have a hidden size of 32.

•**att-RNN [32].** att-RNN employs an attention mechanism that combines textual, visual, and social contextual features. It utilizes Long Short-Term Memory (LSTM) networks to extract textual features and integrates them with visual features through a cross-modal attention mechanism.

•**EANN [20].** The Event Adversarial Neural Networks (EANN) consist of three components: The multimodal feature extractor, the fake news detector, and the event discriminator. The multimodal feature extractor extracts textual and visual features from posts and collaborates with the fake news detector to learn distinctive representations for fake news detection. The event discriminator is responsible for removing specific features related to events. All parameters used in training this model remain consistent with those of the original model.

•**MVAE [21].** This methodology employs an encoding-decoding paradigm to capture shared representations encompassing both visual and textual modalities, to detect fake news. Through the training of a multimodal variational autoencoder, the approach involves the concatenation of text and visual features to derive multimodal representations. These representations undergo decoding, guided by reconstruction loss, to revert them to their original modalities. The resulting multimodal representations are strategically harnessed for the discrimination of fake news.

•**BDANN [36].** Textual features in BDANN are extracted using a pre-trained BERT model, while visual features are obtained through a pre-trained VGG-19 model. Dependency on specific events is mitigated by incorporating a domain classifier.

•**Roberta+CNN [37].** This framework incorporates a dedicated convolutional neural network model for image analysis and a sentence transformer for textual analysis. Features extracted from

visual and textual modalities are embedded through dense layers, ultimately converging to predict deceptive imagery.

•**MEAN [38].** This approach comprises two integral components: A multimodal generator and a dual discriminator. The multimodal generator is instrumental in extracting latent discriminative feature representations for both text and image modalities. For each modality, a decoder is employed to mitigate information loss during the generation process. The dual discriminator consists of a modality discriminator and an event discriminator. These discriminators are designed to classify features based on either modality or event, with network training guided by an adversarial scheme.

This paper employs conventional evaluation metrics for binary classification to assess the model's performance. These metrics include accuracy, precision, recall, and F_1 score. The experimental comparison results are presented in Table 3 and Fig. 6.

Table 3: Comparison of accuracy, precision, recall, and F_1 score for different baselines

			Fake news			Real news		
	Method	Accuracy	Precision	Recall	F_1	Precision	Recall	F_1
Weibo	Textual	0.700	0.680	0.700	0.690	0.710	0.690	0.700
	Visual	0.640	0.580	0.570	0.610	0.640	0.690	0.660
	VQA	0.736	0.797	0.634	0.706	0.695	0.838	0.760
	NeuralTalk	0.726	0.794	0.713	0.692	0.684	0.840	0.754
	att-RNN	0.772	0.854	0.656	0.742	0.720	0.889	0.795
	EANN	0.782	0.827	0.697	0.756	0.752	0.863	0.804
	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
	Roberta+CNN	0.812	0.851	0.784	0.816	0.744	0.826	0.782
	MEAN	**0.894**	**0.900**	0.870	0.890	0.890	**0.910**	**0.900**
	EANBS	0.890	0.870	**0.910**	**0.890**	**0.900**	0.880	0.890
Twitter	Textual	0.630	0.620	0.580	0.600	0.630	0.680	0.650
	Visual	0.590	0.580	0.540	0.560	0.600	0.640	0.620
	VQA	0.631	0.765	0.509	0.611	0.550	0.794	0.650
	NeuralTalk	0.610	0.728	0.504	0.595	0.534	0.752	0.625
	att-RNN	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN	0.648	0.810	0.498	0.617	0.584	0.759	0.660
	MVAE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	BDANN	0.830	0.810	0.630	0.710	0.830	0.930	0.880
	Roberta+CNN	0.853	0.821	**0.943**	**0.877**	**0.913**	0.745	0.820
	MEAN	0.780	0.690	0.840	0.760	0.870	0.740	0.800
	EANBS	**0.860**	**0.850**	0.880	0.860	0.880	**0.840**	**0.860**

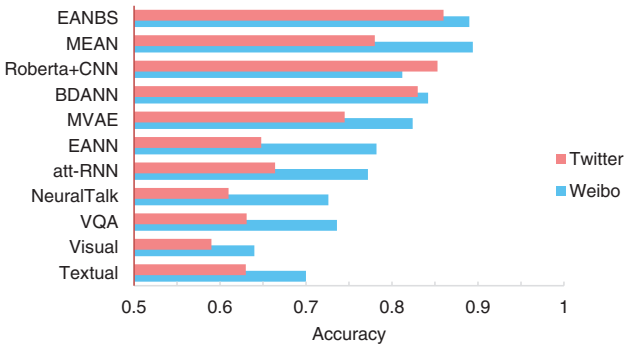


Figure 6: Comparison of experimental results

According to Table 3, on the Weibo dataset, the textual modality demonstrates a significant advantage in the task of detecting fake news compared to the visual modality. Text is more effective in conveying the core content of events, and its embedded semantic information directly facilitates the identification of fake news. In contrast, although the visual modality provides some visual information, its expressive capability is relatively limited, making it challenging to offer as rich semantic information as text. Therefore, in the fake news detection task, text modality proves to be more effective than the visual modality, better distinguishing between genuine and fake news information. While the unimodal model exhibits some effectiveness in fake news detection, their performance remains inadequate compared to multimodal models, further confirming the excellence of multimodal fake news detection methods. Multimodal fake news detection methods can fully leverage diverse information sources, such as text and images, to obtain more comprehensive and enriched feature representations. Text and images complement each other in expressing information, and through the effective fusion of these modalities' features, the model can enhance its ability to identify fake news.

In the realm of multimodal fake news detection models, the MVAE exhibits superior performance by leveraging a multimodal variational autoencoder. Outperforming other models such as VAQ, NeuralTalk, and att-RNN, the EANN and BDANN models introduce Event Adversarial Neural Networks, achieving superior results in fake news detection. Through the incorporation of Event Adversarial Neural Networks, the EANN and BDANN models excel in learning and applying common features among events, thereby demonstrating robust performance in fake news detection tasks. Event Adversarial Networks contribute to the model's ability to learn more general and transferable feature representations, mitigating reliance on specific events and enhancing both robustness and generalization. In comparison to the EANN and BDANN models, MEAN exhibits outstanding performance by learning both modality-invariant and event-invariant features through dual discriminators. On the Twitter dataset, the model's performance is similar to that on the Weibo dataset.

The proposed EANBS fake news detection model demonstrates superior performance across various metrics on both Weibo and Twitter datasets compared to the contrastive models. On the Twitter dataset, the accuracy and precision of fake news detection surpass the best results in the comparative methods by 0.07% and 2.9%. On the Weibo dataset, the recall and precision of fake news detection surpass the best results in the comparative methods by 4% and 1%, respectively. Leveraging similarity representation learning to capture relationships between different modalities, the model comprehensively understands news content, thereby improving the accuracy and robustness of fake news detection. The introduction of adversarial networks further enhances the model's performance,

aiding in learning more robust feature representations and eliminating dependence on specific events. This improvement increases the model's generalization capability towards unseen events, effectively identifying fake news and enhancing the overall quality of news on social media.

4.5 Analysis of Ablation Experiments

To verify the significance of the model components outlined in this paper, we created several model variants. These variants primarily fall into three types: EANBS-SIM, which eliminates similarity representation learning and reasoning; EANBS-BERT, which removes the BERT component; and EANBS-GAN, which excludes adversarial neural networks.

(1) Eliminate Similarity Representation Learning and Reasoning. The textual modality employs Text-CNN and BERT models to extract textual features, while the pre-trained VGG-19 model is used to extract visual features for the visual modality. Subsequently, the resulting feature vectors from both modalities are merged and used as input to both the event classifier and fake news detector.

(2) Eliminate BERT Component. The Text-CNN model is employed for the textual modality, while the pre-trained VGG-19 model is used for the visual modality. The similarity representation learning and reasoning module calculates the similarity between the extracted textual and visual features. Finally, textual features, visual features, and their similarity are concatenated as input for both the event classifier and the fake news detector.

(3) Eliminates the Adversarial Neural Networks. For the textual modality, both Text-CNN and BERT models are employed to extract textual features, while the pre-trained VGG-19 model is used for visual feature extraction. Subsequently, the similarity representation learning and reasoning module computes the similarity between the textual features obtained from Text-CNN and the visual features obtained from the pre-trained VGG-19. The resulting textual features, visual features, and their similarity are then combined and input into a fully connected layer with SoftMax for classification.

The results of ablation experiments are presented in Table 4, indicating that the removal of any component of the model results in a noticeable decrease in classification accuracy. This underscores the effectiveness of each model component in the experiments. The introduction of BERT provides our model with enhanced semantic understanding and more robust feature representation capabilities, aiding in capturing crucial features within fake news text. Utilizing the BERT model allows for more accurate comprehension of semantic relationships in text, enabling the identification of hidden information and effective discrimination between fake and genuine news. The similarity representation learning and inference model proves effective in capturing the relationship, i.e., similarity, between news text and image information in fake news detection. Through this model, common features between text and images are learned, facilitating inference regarding their degree of similarity. Such a model enables a more comprehensive understanding of news events, leading to more accurate assessments of news authenticity.

While the role of adversarial neural network models in fake news detection may be relatively weaker, comprehensive experiments combining BERT, similarity representation learning and inference models, and adversarial neural network models demonstrate superior performance compared to the ablation experiment results. This suggests the crucial complementary and synergistic effects among these models in enhancing fake news detection. By synergistically leveraging BERT's semantic understanding, the association-capturing capabilities of similarity representation learning and inference models, and the feature fusion optimization of adversarial neural network models, we achieve outstanding results in fake news detection on two datasets. This multimodal combination approach

offers new perspectives and technical means for fake news detection research, enhancing model performance and reliability.

Table 4: Comparison of results of ablation experiments

			Fake news			Real news		
	Method	Accuracy	Precision	Recall	F_1	Precision	Recall	F_1
Weibo	**EANBS**	**0.890**	**0.870**	**0.910**	**0.890**	**0.900**	**0.880**	**0.890**
	EANBS-SIM	0.870	0.860	0.880	0.870	0.880	0.870	0.870
	EANBS-BERT	0.800	0.780	0.810	0.800	0.820	0.780	0.800
	EANBS-GAN	0.880	0.860	0.890	0.880	0.900	0.860	0.880
Twitter	**EANBS**	**0.860**	**0.850**	**0.880**	**0.860**	**0.880**	0.840	**0.860**
	EANBS-SIM	0.810	0.780	0.850	0.810	0.850	0.780	0.810
	EANBS-BERT	0.780	0.790	0.750	0.770	0.770	0.810	0.790
	EANBS-GAN	0.830	0.900	0.790	0.840	0.760	**0.880**	0.820

4.6 Visualization Analysis

To further assess the efficacy of the event classifier, we visualize the textual feature representations obtained by the model in this paper, both with and without adversarial neural networks, using the Weibo test dataset. Fig. 7 depicts the visualization of textual representations, where red dots represent labeled features of fake news, and blue dots represent labeled features of real news. Based on the feature distribution, it can be observed that the model without the adversarial neural networks can learn distinguishable features. However, the learned features are still intertwined when compared to the feature representations learned by the model in this paper. This also indicates that during the training phase, the event classifier endeavors to eliminate dependencies between feature representations and specific events. Through the minimax game, the multimodal feature extractor can acquire invariant feature representations of different events, enabling the learned generic features to be employed for transfer learning to discern the authenticity of breaking fake news in sudden events. This enhances the model's transferability and its ability to generalize to new events, ultimately improving the performance of fake news detection.

Figure 7: Visualization of textual feature representations learned on the Weibo test dataset

4.7 Parameter Analysis

To examine the influence of model parameters on performance, we conducted several experimental sets on two datasets, Weibo and Twitter, followed by a comparative analysis of their outcomes.

These experiments primarily focused on variations in the learning rate and different settings of the optimizer. Figs. 8 and 9 demonstrate the comparison of the experimental results.

(1) Effect of Learning Rate: Fig. 8 illustrates the impact of various learning rate values on the performance of the proposed model on two datasets, Weibo and Twitter. As shown in the figure, the model achieved its highest accuracy and F_1 score on two datasets when the learning rate was set to 0.001. Therefore, in our experiments, we chose to set the learning rate to 0.001.

(2) Impact of Optimizers: Fig. 9 illustrates the impact of different optimizers on the performance of the proposed model on the microblogging dataset. In terms of model prediction accuracy, the use of the SGD optimizer results in a classification accuracy of around 0.85 after 50 rounds of training, while the Adam optimizer achieves a classification accuracy close to 1.0. Therefore, in our experiments, we opted to use the Adam optimizer for model training. The optimizer guides the various parameters of the loss function during the backpropagation process to update in the correct direction with an appropriate magnitude, continuously approaching the global minimum. In deep learning, selecting an appropriate optimizer can significantly improve both training efficiency and accuracy.

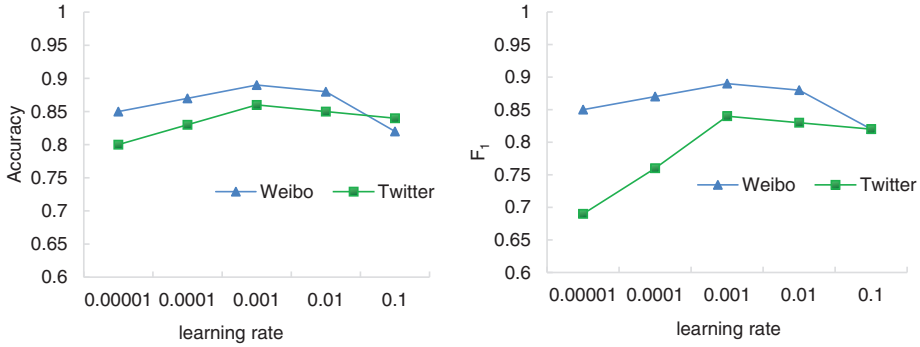


Figure 8: Impact of various learning rates on results

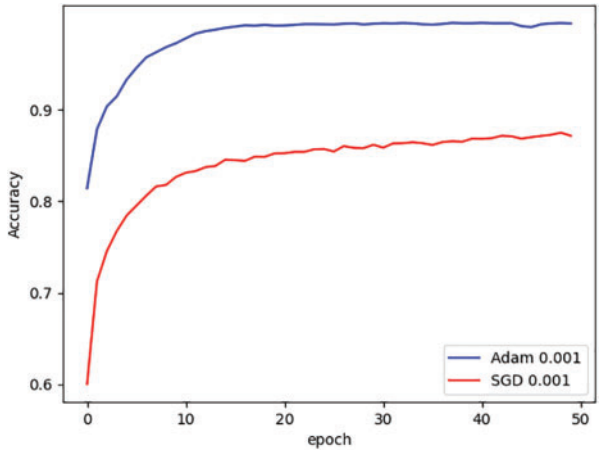
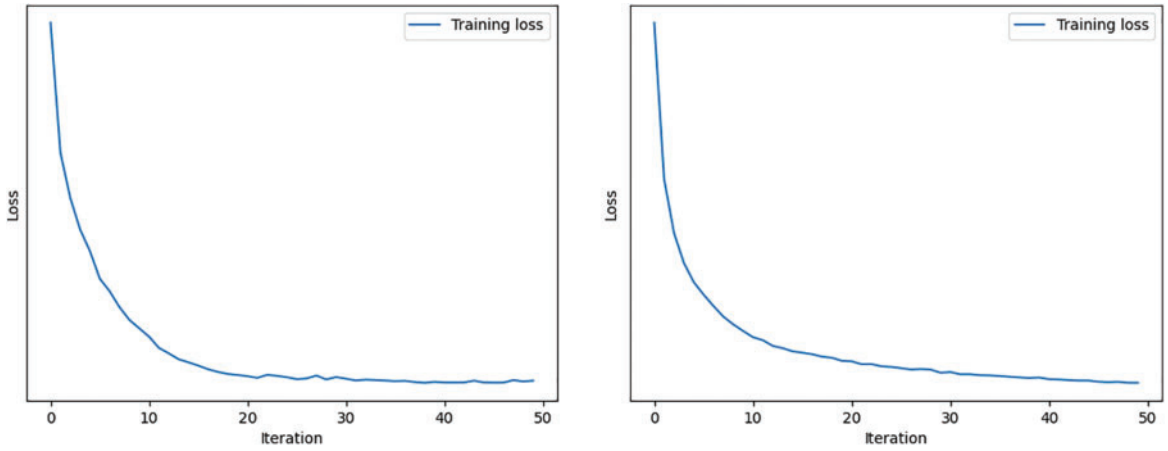


Figure 9: Effects of different optimizers on the results

4.8 Convergence Analysis

To verify the convergence of the proposed model, we selected parameters $\alpha = 10$ and $\beta = 0.75$ for the experiments. Fig. 10 illustrates the changes in the loss function on two datasets, Weibo and Twitter. In the initial stages of training, the loss rapidly decreases, indicating that the choice of learning rate is appropriate, and the model has entered the gradient descent process. As training progresses, the loss function gradually stabilizes, indicating that the model has reached a certain state of equilibrium, demonstrating that the model has undergone effective learning.



(a) Weibo (b) Twitter

Figure 10: Change of loss

4.9 Fault Case Study

To further illustrate the performance of our proposed method, we collected and analyzed some failure cases. Figs. 11a and 11b depict two instances where our proposed approach failed to detect fake news. In Fig. 11a, the excessive use of exclamation marks in the post's text, along with a presentation format not typical of traditional news media, is likely to impact the model's discriminative ability on the data. In Fig. 11b, the post's text is exceptionally short, resulting in suboptimal performance of the proposed similarity reasoning and adversarial network models. Additionally, to explore enhanced performance, we intend to address these limitations in future work.

(a) #GuangmingLandslide#It's happening quickly!! Just 5 meters away from a miracle of life! According to on-site reporters, the image now shows the fourth floor, and signs of life are on the first or second floor!! Go, Zhongshan Fire Rescue Team!!! (Reporter: Jiang Tianli)

(b) Oh, My god!

Figure 11: Certain fake news that cannot be correctly classified by the proposed EANBS

5 Conclusion

The spread of fake news not only damages the reputation of news organizations but also pollutes the online information landscape, posing a significant threat to the growth of social media. This paper aims to address the issue of detecting fake news in social media by constructing a multimodal social media detection model that utilizes similarity reasoning and adversarial networks. This is achieved by examining the similarities between the textual and visual components of fake news content. Through the analysis of the results from various feature experiments, we have discovered that the similarity features between the textual and visual aspects are highly effective in distinguishing between fake and real news. By implementing a game between the feature extractor and classifier, the model can acquire event-invariant representations by eliminating specific event-related features, thereby reducing the strong dependence on specific events in fake news. Moreover, the model can detect shared event characteristics in fake news, which improves its ability for feature transfer and the identification of fake news in emerging events. We have conducted numerous experiments on two datasets to demonstrate the model's effectiveness. However, it is worth noting that we have primarily focused on the correlation between text and image in our model. As a result, we intend to explore the integration of video information features in our future research.

Acknowledgement: The authors would like to thank the anonymous referees for their valuable comments and helpful suggestions.

Funding Statement: This paper is supported by the National Natural Science Foundation of China (No. 62302540), with author F.F.S. For more information, please visit their website at https://www.nsfc.gov.cn/. Additionally, it is also funded by the Open Foundation of Henan Key Laboratory of Cyberspace Situation Awareness (No. HNTS2022020), where F.F.S is an author. Further details can be found at http://xt.hnkjt.gov.cn/data/pingtai/. The research is also supported by the Natural Science Foundation of Henan Province Youth Science Fund Project (No. 232300420422), and for more information, you can visit https://kjt.henan.gov.cn/2022/09-02/2599082.html. Lastly, it receives funding from the Natural Science Foundation of Zhongyuan University of Technology (No. K2023QN018), where F.F.S is an author. You can find more information at https://www.zut.edu.cn/.

Author Contributions: The authors confirm contribution to the paper as follows: Study conception and design: Fangfang Shan, Huifang Sun; data collection: Huifang Sun, Mengyi Wang; analysis and interpretation of results: Huifang Sun; draft manuscript preparation: Huifang Sun. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The Weibo data used to support the findings of this study have been deposited on the website: https://drive.google.com/file/d/14VQ7EWPiFeGzxp3XC2DeEHi-BEisDINn/view?usp=sharing. The Twitter data used to support the findings of this study have been deposited on the website: https://github.com/MKLab-ITI/image-verification-corpus.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] J. Wang, Y. Liu, and S. J. Ji, "Matrix factorized bilinear pooling for multi-modal fusion fake news detection (In Chinese)," *Appl. Res. Comput.*, vol. 39, no. 10, pp. 2968–2973+2978, 2022.

[2] P. Qi, J. Cao, and Q. Sheng, "Semantics-enhanced multi—modal fake news detection," *J. Comput. Res. Dev.*, vol. 58, no. 7, pp. 1456–1465, 2021.
[3] J. Z. Zhang, J. Cao, and D. Radcliffe, "The impact of COVID-19 on journalism in emerging economies and the global south," *Youth Journalist*, vol. 699, no. 7, pp. 99–100, 2021.
[4] K. Shu, A. Sliva, S. Wang, J. L. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explor. Newsletter*, vol. 19, no. 1, pp. 22–36, 2017. doi: 10.1145/3137597.3137600.
[5] O. Ajao, D. Bhowmik, and S. Zargari, "Sentiment aware fake news detection on online social networks," in *Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, UK, 2019, pp. 2507–2511.
[6] M. A. Alonso, D. Vilares, C. Gómez-Rodríguez, and J. Vilares, "Sentiment analysis for fake news detection," *Electron.*, vol. 10, no. 11, pp. 1348, 2021. doi: 10.3390/electronics10111348.
[7] N. Islam *et al.*, "Ternion: An autonomous model for fake news detection," *Appl. Sci.*, vol. 11, no. 19, pp. 9292, 2021. doi: 10.3390/app11199292.
[8] X. C. Liu, "Research on fake news detection based on machine learning algorithms," *Inf. Technol. Inf.*, no. 9, pp. 237–239, 2021.
[9] J. Ma *et al.*, "Detecting rumors from microblogs with recurrent neural networks," in *Twenty-Fifth Int. Joint Conf. Artif. Intell.*, New York, NY, USA, 2016, pp. 3818–3824.
[10] F. Yu, Q. Liu, S. Wu, L. Wang, and T. Tan, "A convolutional approach for misinformation identification," in *Proc. Twenty-Sixth Int. Joint Conf. Artif. Intell.*, Melbourne, Australia, 2017, pp. 3901–3907.
[11] J. Ma, W. Gao, and K. F. Wong, "Detect rumors on twitter by promoting information campaigns with generative adversarial learning," in *The World Wide Web Conf.*, San Francisco, USA, 2019, pp. 3049–3055.
[12] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel visual and statistical image features for microblogs news verification," *IEEE Trans. Multim.*, vol. 19, no. 3, pp. 598–608, 2016. doi: 10.1109/TMM.2016.2617078.
[13] T. Mahmood, Z. Mehmood, M. Shah, and T. Saba, "A robust technique for copy-move forgery detection and localization in digital images via stationary wavelet and discrete cosine transform," *J. Vis. Commun. Image Rep.*, vol. 53, no. 5, pp. 202–214, 2018. doi: 10.1016/j.jvcir.2018.03.015.
[14] S. Farooq, M. H. Yousaf, and F. Hussain, "A generic passive image forgery detection scheme using local binary pattern with rich models," *Computers & Electrical Engineering*, vol. 62, pp. 459–472, 2017.
[15] A. Peng, Y. Wu, and X. Kang, "Revealing traces of image resampling and resampling antiforensics," *Adv. Multimed.*, vol. 2017, pp. 7130491, 2017. doi: 10.1155/2017/7130491.
[16] J. Zeng, S. Tan, B. Li, and J. Huang, "Large-scale JPEG image steganalysis using hybrid deep-learning framework," *IEEE Trans. Inf. Forens. Secur.*, vol. 13, no. 5, pp. 1200–1214, 2017. doi: 10.1109/TIFS.2017.2779446.
[17] W. Quan, K. Wang, D. M. Yan, and X. Zhang, "Distinguishing between natural and computer-generated images using convolutional neural networks," *IEEE Trans. Inf. Forens. Secur.*, vol. 13, no. 11, pp. 2772–2787, 2018. doi: 10.1109/TIFS.2018.2834147.
[18] Z. Zhao *et al.*, "Social-aware movie recommendation via multimodal network learning," *IEEE Trans. Multimed.*, vol. 20, no. 2, pp. 430–440, 2017. doi: 10.1109/TMM.2017.2740022.
[19] Z. W. Jin, J. Cao, B. Wang, R. Wang, and Y. D. Zhang, "Rumor detection on social media with multimodal feature fusion," *J. Nanjing Univ. Inf. Sci. Technol.*, vol. 9, no. 6, pp. 583–592, 2017.
[20] Y. Q. Wang *et al.*, "EANN: Event adversarial neural networks for multi-modal fake news detection," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, London, UK, 2018, pp. 849–857.
[21] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "MVAE: Multimodal variational autoencoder for fake news detection," in *The World Wide Web Conf.*, San Francisco, CA, USA, 2019, pp. 2915–2921.
[22] L. Cui, S. Wang, and D. Lee, "SAME: Sentiment-aware multi-modal embedding for detecting fake news," in *Proc. 2019 IEEE/ACM Int. Conf. Adv. Soc. Netw. Anal. Min.*, Vancouver, Canada, 2019, pp. 41–48.
[23] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. I. Satoh, "Spotfake: A multi-modal framework for fake news detection," in *2019 IEEE Fifth Int. Conf. Multimed. Big Data*, Changsha, China, 2018, pp. 39–47.

[24] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.
[25] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conf. Comput. Vis. Pattern Recogn.*, Miami, Florida, USA, 2009, pp. 248–255.
[26] S. Singhal, A. Kabra, M. Sharma, R. R. Shah, T. Chakraborty and P. Kumaraguru, "SpotFake+: A multimodal framework for fake news detection via transfer learning (student abstract)," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 10, pp. 13915–13916, 2020. doi: 10.1609/aaai.v34i10.7230.
[27] Z. L. Yang, Z. H. Dai, Y. M. Yang, J. Carbonell, R. Salakhutdinov and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2019, pp. 5753–5763.
[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones and A. N. Gomez, "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, vol. 30, 2017, pp. 5998–6008.
[29] H. Diao, Y. Zhang, L. Ma, and H. Lu, "Similarity reasoning and filtration for image-text matching," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 2, pp. 1218–1226, 2021. doi: 10.1609/aaai.v35i2.16209.
[30] K. H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 201–216.
[31] Z. Kuang *et al.*, "Fashion retrieval via graph reasoning networks on a similarity pyramid," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Seoul, South Korea, 2019, pp. 3066–3075.
[32] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proc. 25th ACM Int. Conf. Multimed.*, New York, USA, 2017, pp. 795–816.
[33] C. Boididou *et al.*, "Verifying multimedia use at mediaeval 2015," in *MediaEval 2015 Workshop*, Wurzen, Germany, 2015.
[34] S. Antol *et al.*, "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 2425–2433.
[35] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Boston, USA, 2015, pp. 3156–3164.
[36] T. Zhang *et al.*, "BDANN: BERT-based domain adaptation neural network for multi-modal fake news detection," in *2020 Int. Joint Conf. Neural Netw. (IJCNN)*, Glasgow, UK, 2020, pp. 1–8.
[37] B. Singh and D. K. Sharma, "Predicting image credibility in fake news over social media using multi-modal approach," *Neural Comput. Appl.*, vol. 34, no. 24, pp. 21503–21517, 2022. doi: 10.1007/s00521-021-06086-4.
[38] P. Wei, F. Wu, Y. Sun, H. Zhou, and X. Y. Jing, "Modality and event adversarial networks for multi-modal fake news detection," *IEEE Signal Process. Letters*, vol. 29, no. 2, pp. 1382–1386, 2022. doi: 10.1109/LSP.2022.3181893.