Wu et al. Cybersecurity (2026) 9:155
https://doi.org/10.1186/s42400-026-00550-1

Cybersecurity

RESEARCH Open Access

Fake news detection with GAN-augmented contrastive learning and multimodal attention

Cong Wu[1*], Jing Chen[1], Yebo Feng[2], Ju Jia[3], Tingting Xu[4], Zijian Zhang[5], Jiahua Xu[6,7], Teng Li[8] and Yang Liu[2]

Abstract

The rapid proliferation of fake news in digital media has emerged as a major threat to information credibility and public trust. Although recent advances have explored multimodal learning for fake news detection, existing models often fail to effectively integrate heterogeneous data sources and remain vulnerable to adversarial manipulations. To address these challenges, we propose MADSL (Multimodal Adversarial Deep Semantic Learning), a robust multimodal fake news detection framework that unifies generative adversarial networks (GANs) with supervised contrastive learning. Specifically, MADSL employs a multi-layer joint attention mechanism to align and fuse textual and visual features, while adversarial training encourages the extraction of event-invariant representations, enhancing generalizability across unseen news events. Additionally, contrastive learning with adversarial perturbations further strengthens feature discrimination and robustness against attacks. Extensive experiments on benchmark Twitter and Weibo datasets demonstrate that MADSL achieves state-of-the-art accuracy (85.3%) and maintains stable performance with only a 1.1% drop under adversarial conditions, outperforming existing methods in both detection accuracy and resilience. These results underscore MADSL 's effectiveness in advancing robust multimodal fake news detection and promoting digital information integrity.

Keywords Fake news detection, Contrastive learning, Generative adversarial networks

Introduction

The rapid expansion of digital platforms has significantly exacerbated the spread and impact of fake news (Choi et al. 2017; Qu et al. 2020; Capuano et al. 2023; Bo et al. 2025; Iqbal et al. 2025; Liu et al. 2018), challenging the integrity of information and undermining the quality of public discourse. In today's digital age, where social media and online news are prevalent, fake news quickly influences public opinion, with the potential to affect election outcomes and incite unrest. A striking statistic reveals that 86% of online global citizens have encountered fake news, highlighting its widespread influence. This trend emphasizes the urgent need for more sophisticated and nuanced fake news detection tools. Traditional content verification methods, which often rely on single modalities such as text, images, or videos, are increasingly inadequate due to the complexity of digital misinformation. There is a critical demand for advanced strategies that can effectively distinguish truth from falsehood, essential not only for preserving the integrity of information but also for supporting the foundational principles of informed and democratic societies (Zhang et al. 2022). Ensuring the development and deployment of effective fake news detection tools is paramount in maintaining factual and trustworthy communication in the fight against misinformation.

Recent advancements in deep learning have dramatically reshaped the landscape of fake news detection.

*Correspondence:
Cong Wu
cancwu@whu.edu.cn
[1] Wuhan University, Wuhan, China
[2] Nanyang Technological University, Singapore, Singapore
[3] Southeast University, Nanjing, China
[4] City University of Macau, Macau, China
[5] Beijing Institute of Technology, Beijing, China
[6] University College London, London, UK
[7] Exponential Science Foundation, Lugano, Switzerland
[8] Xidian University, Xi'an, China

Existing efforts increasingly employ neural network architectures such as Convolutional Neural Networks (CNNs) and use Recursive Neural Networks (RNNs) to extract and analyze features from both textual and visual content. This evolution is further enhanced by the adoption of sophisticated Natural Language Processing (NLP) techniques, including state-of-the-art models like BERT (Devlin et al. 2019) and GPT (Brown et al. 2020), which provide a nuanced analysis of language crucial for detecting subtle cues in fake news narratives (Yang et al. 2021; Lin et al. 2023). These developments have significantly improved our ability to process and understand complex content structures, resulting in more accurate detection methods. However, despite these technological strides, notable research gaps in the field still persist.

(i) Adaptability to new events: Current fake news detection models struggle to adapt to new and rapidly evolving events. Although they are good at feature extraction, they often fail to generalize effectively to unfamiliar situations (Zhang and Ghorbani 2020; Zhang et al. 2023; Yin et al. 2024; Zhang et al. 2024). This is due to their reliance on event-specific features, which are ineffective in new scenarios. Therefore, there is a need to develop models capable of extracting event-invariant features to maintain accuracy across diverse events.

(ii) Robustness against adversarial attacks: Improving the robustness and generalization of models to withstand adversarial attacks is a significant challenge (Wang et al. 2018; Zhang et al. 2024; Kasampalis et al. 2024; Wang et al. 2023). Adversarial attacks exploit model weaknesses and often mimic legitimate data. The rise of sophisticated generative language models increases the risk of these attacks (Ali et al. 2021; Wang et al. 2018). Current methods frequently fail against such perturbations. Thus, integrating adversarial training to enhance model resilience is crucial for reliable fake news detection, given its significant societal impact.

Our approach. In this paper, we introduce Multimodal Adversarial Deep Semantic Learning, namely MADSL, a novel approach for enhancing fake news detection and generalization. MADSL leverages multimodal feature extraction by integrating contrastive learning and adversarial training to handle the complexities of fake news. The system combines Text-CNN (Chen 2015) for extracting textual features and VGG-19 (Simonyan and Zisserman 2015) for analyzing visual content, facilitating a comprehensive analysis of both text and images. A multilayer joint attention mechanism aligns textual and visual features, enhancing the model's ability to identify subtle indicators of manipulated content. This fusion of multimodal inputs and sophisticated processing techniques significantly improves MADSL 's adaptability to new and evolving fake news scenarios, making it a robust tool against misinformation.

To address the challenge of rapidly responding to emerging news events, MADSL incorporates a GAN model in conjunction with contrastive learning. This strategy enhances the system's capacity to generate event-invariant features, thereby bolstering its generalization capabilities across various scenarios. MADSL utilizes supervised contrastive learning to effectively differentiate between news categories, fostering the development of robust features crucial for accurate classification. The model's resilience is further enhanced by the introduction of gradient-based adversarial perturbations through Projected Gradient Descent (PGD) (Madry et al. 2017; Jia et al. 2025c) during the training process, strengthening its defenses against adversarial attacks (Jing et al. 2023).

- We propose MADSL, which innovatively combines GANs with adversarial contrastive learning. This integration facilitates the generation of robust, event-independent features that significantly improve the model's adaptability and generalization capabilities across diverse and novel events.
- MADSL employs GANs for enhancing feature extraction, creating an adversarial dynamic between the multimodal feature extractor and an event classifier to generate event-independent, robust features.
- The model incorporates supervised contrastive learning, effectively differentiating between real and fake news by analyzing feature representation similarities and disparities, enhancing its discernment and generalization.
- We rigorously evaluated MADSL's performance on two datasets. The results demonstrate its superior accuracy of 85.3% in fake news detection and exceptional resilience against adversarial attacks, setting a new benchmark in the field.

Section 2 presents the research background and Section 3 details the system design of MADSL. Section 4 describes the experimental setup and reports the performance of MADSL. Section 5 discusses the experimental results and analysis. Finally, Section 6 concludes the paper and outlines future research directions.

Background

In this section, we introduce the background of text feature extraction, attention mechanism, and contrastive learning.

Text feature extraction

Word embedding techniques have transformed natural language processing by providing nuanced

representations of words or phrases as vectors within a compact vector space, which significantly surpasses the simplistic and dimensionally excessive one-hot encoding (Mikolov et al. 2013). By capturing semantic correlations and contextual meanings, word embeddings facilitate a deeper understanding of language. They have become fundamental in various NLP applications, from text classification to sentiment analysis, thanks to their ability to encode intricate linguistic attributes. Notably, architectures like Word2Vec (Mikolov et al. 2013) and GloVe (Pennington et al. 2014) have been crucial to these developments. Word2Vec enhances context prediction with its CBOW and Skip-Gram models, while GloVe combines matrix factorization with context window techniques to effectively utilize global word co-occurrence patterns, establishing a solid foundation for sophisticated computational models to analyze and interpret language deeply.

CNNs have become a cornerstone in deep learning, renowned for their proficiency in pattern recognition within visual data (Gao et al. 2024). Characterized by their convolutional layers, which employ learnable filters to extract local feature representations like edges and textures, CNNs are structured with convolutional, pooling, and fully connected layers that facilitate a layered learning approach. This architecture allows CNNs to progressively capture more complex and abstract features, making them exceptionally effective in various computer vision tasks. Beyond visual analysis, CNNs also play a significant role in NLP, where they analyze text to identify and extract critical features such as n-grams, effectively recognizing local sequential dependencies in textual data. This ability enables CNNs to discern phraseology and word clusters that signify underlying sentiments or classifications, making them invaluable for tasks ranging from text classification to sentiment analysis. By processing word embeddings through their convolutional layers to distill semantic and syntactic patterns, CNNs provide a nuanced and comprehensive textual representation, demonstrating their versatility and effectiveness across both visual and linguistic domains.

Attention mechanism

Attention mechanisms in deep learning, inspired by human cognitive focus, have significantly enhanced the performance of models across a broad spectrum of tasks (Vaswani et al. 2017). Central to these mechanisms is their ability to selectively concentrate on relevant parts of the input data, effectively mimicking human attention processes. This selective focus dynamically assigns weights to different input segments, allowing models to emphasize important features while disregarding irrelevant ones. In NLP, attention mechanisms excel at highlighting crucial words or phrases for tasks like sentiment analysis or machine translation, focusing on the most informative elements to enhance prediction accuracy and contextual relevance.

Moreover, attention mechanisms address the constraints of traditional neural network architectures in sequence-to-sequence tasks, where conventional models often reduce entire input sequences to fixed-dimensionality vectors, potentially losing vital context, particularly in lengthy sequences. By iteratively focusing on specific segments of the input throughout the generation of outputs, attention mechanisms maintain the relevance of the context. This method significantly improves the model's capacity to manage long-range dependencies and understand complex data structures more effectively. As a result, attention mechanisms have become fundamental in advancing deep learning, markedly improving accuracy and deepening the contextual understanding in applications ranging from language translation to image captioning and speech recognition.

Contrastive learning

Contrastive learning has emerged as a fundamental technique in self-supervised learning, widely applied across both Computer Vision and Natural Language Processing fields (Jia et al. 2025a; Wu et al. 2025a, b). A notable model, SimCLR (Chen et al. 2020), has significantly advanced the field of image representation learning by adopting a contrastive method. The core principle of SimCLR involves minimizing the distance between various augmented versions of the same image within the embedding space, while simultaneously maximizing the distance between different images. This approach effectively reduces reliance on the extensive, annotated datasets typically required in supervised learning frameworks, offering a more efficient and scalable alternative.

This learning model uses internally generated labels for supervision, ensuring that the developed representations are versatile across various downstream tasks. However, a significant limitation of the self-supervised contrastive learning approach is its oversight in recognizing the interconnectedness of features among images from the same class, potentially resulting in disjointed feature space representations for similar categories. To address this, supervised contrastive learning has been introduced, leveraging explicit label information to align representations of the same class more closely within the embedding space, enhancing classification accuracy.

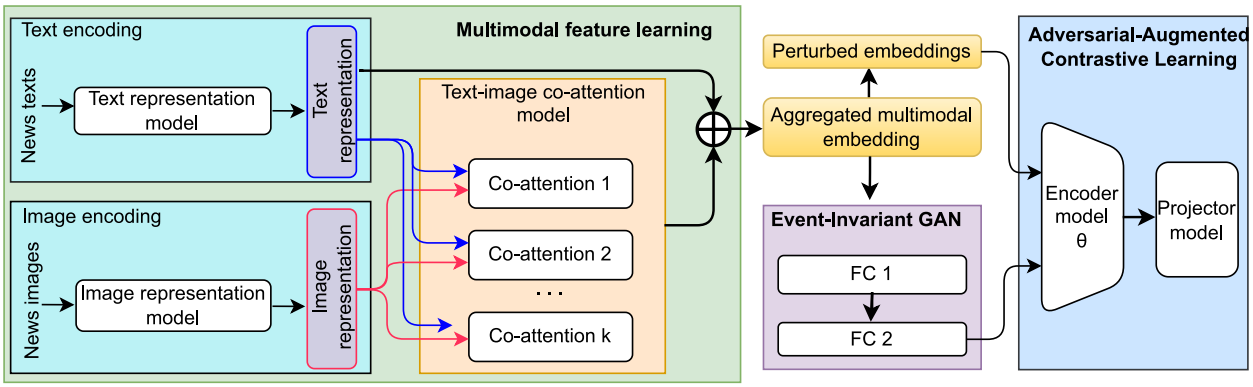

Fig. 1 Overview of MADSL

System design

In this section, we first introduce the system overview, then illustrate details of system design.

System overview

MADSL consists of three integrated stages: Multimodal Feature Learning, Event-Invariant Generative Adversarial Learning, and Adversarial-Augmented Contrastive Learning. The first stage harnesses text and image encoding techniques coupled with a multi-layer attention mechanism to extract and synthesize high-dimensional semantic features. The second stage employs a GAN to enhance the model's adaptability and generalization across diverse and unforeseen event contexts by training the feature extractor and event classifier in an adversarial setup. The final stage incorporates supervised contrastive learning with adversarial perturbations, significantly boosting the model's resilience against sophisticated adversarial attacks and enhancing its discriminative capabilities. Together, these stages enable MADSL to effectively discern genuine from fake content, ensuring robust performance in real-world scenarios and establishing it as a potent tool against misinformation.

In our model for fake news detection, as illustrated in Fig. 1, we integrate three key components: multimodal feature extraction, event-invariant generative adversarial learning, and adversarial-augmented contrastive learning. The first stage, multimodal feature extraction, employs a Text Encoder using a CNN for semantic analysis of text and an Image Encoder with VGG-19 for visual feature extraction. These modalities are then seamlessly fused via a sophisticated multi-layer attention mechanism, enhancing the model's capacity to discern and synthesize crucial aspects of both textual and visual information.

The second stage involves event-invariant generative adversarial learning, utilizing a GAN to cultivate the model's adaptability to diverse events and resilience against the evolving nature of fake news. This is achieved through a min-max game between the generator, which produces modified examples, and the discriminator, which learns to differentiate between real and altered features. Complementing this, the model employs adversarial-augmented contrastive learning to refine its understanding of authentic versus deceptive content. This approach clusters genuine news features closely in the embedding space, facilitating accurate discrimination. The model's training is meticulously conducted using Stochastic Gradient Descent (SGD) with mini-batch optimization, ensuring a stable and effective learning process capable of robust fake news detection (Fig. 2).

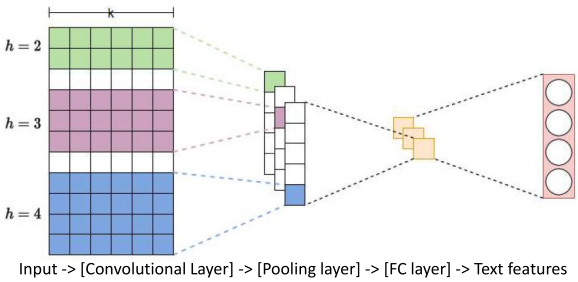

Fig. 2 Structure of CNN model for text feature extraction

Multimodal feature learning

The first stage is multimodal feature learning (Zhang et al. 2025), consisting of text encoding, image encoding, and text-image co-attention. This process aims to derive high-dimensional semantic features from textual and visual content.

Text encoding. Extracting high-dimensional semantic features from text is crucial for fake news detection.

Table 1 Structure of CNN model

Layer	# Input dims	# Channels	# Kernel size	Size
Convolution layer	256*32	4*16	{1,2,3,4} *32	256*64
Pooling layer	256*64	1	256	64
Fully connected (FC) layer	64	1	-	32

Automated methods analyze various aspects of news content, such as article text, sources, headlines, and associated media, using features like sentence segmentation, tokenization, part-of-speech tagging, lexical properties, bag-of-words, term frequency, and syntax to identify deceptive cues and writing styles effectively.

Our text representation model, inspired by the TextCNN architecture, comprises four main components: an input/Word2Vec layer, a convolutional layer, a pooling layer, and a fully connected layer. Text data is converted into a matrix $X \in \mathbb{R}^{256\times32}$, where each row represents a word embedded into a 32-dimensional vector using Word2Vec. This layer uses multiple kernel sizes to extract local n-gram features. For kernel sizes $k \in \{1,2,3,4\}$, each with 16 filters, the feature maps H_k are computed as:

$$H_k = \text{ReLU}(W_k * X + b_k), \tag{1}$$

where W_k is the filter for size k, and $*$ denotes the convolution operation. Max pooling is applied to each feature map to highlight the most significant features and reduce dimensionality. For a feature map H_k of size 256×64:

$$P_k = \text{max\_pool}(H_k, \text{kernel\_size} = 256), \tag{2}$$

resulting in a 64-dimensional vector for each k (Table 1). The pooled features P_k are concatenated and passed through a fully connected layer to form a 32-dimensional feature vector f:

$$f = \sigma(W_f \cdot [P_1; P_2; P_3; P_4] + b_f), \tag{3}$$

where σ is an activation function (e.g., ReLU or Tanh), W_f is the weight matrix, and b_f is the bias term. Regularization techniques like Dropout or L2 regularization are applied to prevent overfitting and improve generalization.

This structure effectively captures and integrates textual features, making it well-suited for subsequent fusion with image features in fake news detection.

Image Encoding. In MADSL, image encoding leverages the VGG-19 neural network, renowned for its depth and ability to extract detailed visual information. Developed by the Visual Geometry Group at Oxford University, VGG-19 features a streamlined architecture with small convolutional filters that enhance depth and performance. The network comprises 16 convolutional layers, interspersed with max pooling layers, and 3 fully connected layers, culminating in a softmax layer for classification tasks.

The VGG-19 network is chosen for its efficient architecture, which uses multiple small-sized convolutional kernels to increase non-linearity and reduce the number of parameters. Although computationally demanding due to its fully connected layers, modifications such as removing these layers can maintain performance while reducing overhead.

In our model, VGG-19's robust application in image classification, transfer learning, and feature extraction makes it particularly suitable. The network processes input images through its 19 layers to produce a visual feature representation $R_{V_{vgg}}$. To ensure compatibility between image and text feature vectors, an additional fully connected layer is added after VGG-19's last layer. During joint training with the text feature extraction network, the pre-trained VGG-19 parameters are kept static to avoid overfitting. The resulting p-dimensional image feature representation $R_V \in \mathbb{R}^p$ is formulated as follows:

$$R_V = \sigma(W_{vf} \cdot R_{V_{vgg}} + b_{vf}), \tag{4}$$

where $R_{V_{vgg}}$ is the visual feature representation from the pre-trained VGG-19 network, W_{vf} is the weight matrix of the fully connected layer within the visual feature extractor, b_{vf} is the bias term, and σ is an activation function (e.g., ReLU) that introduces non-linearity to the feature representation. This integration allows the model to effectively combine detailed visual information with textual features, enhancing its ability to detect fake news.

Text-image co-attention. Recognizing the correlation between the semantic content of text and specific regions in images, we have developed a multi-layer joint attention mechanism to selectively align text and image features, enhancing multimodal fusion effectiveness by minimizing noise from irrelevant data. Using the text vector R_T and image vector R_V, we input these into a single-layer neural network to generate an attention distribution over the image regions through the softmax function, enabling precise alignment of relevant features.

$$\begin{aligned} h_A &= \tanh(W_{V,A}R_V \oplus (W_{T,A}R_T + b_A)), \\ p_V &= \text{softmax}(W_P h_A + b_P), \end{aligned} \tag{5}$$

where $R_V \in \mathbb{R}^{d\times m}$, with d representing the dimension of the image feature and m the number of image regions, and $R_T \in \mathbb{R}^d$ a d-dimensional vector. Let $W_{V,A}, W_{T,A} \in \mathbb{R}^{k\times d}, W_P \in \mathbb{R}^{1\times k}$, then $p_V \in \mathbb{R}^m$ is an m-dimensional vector, corresponding to the attention probability of each image region given R_T. The operation

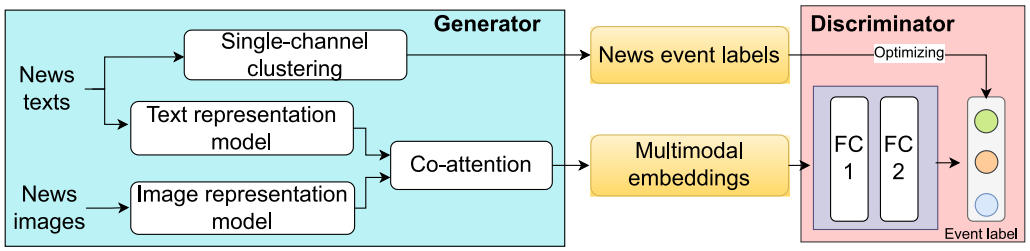

Fig. 3 Event-invariant GAN-based learning

$\oplus$ denotes the addition of a matrix and a vector, where the vector is broadcast to each column of the matrix.

Based on the attention distribution obtained, we now calculate the weighted sum of the image vectors, each coming from a region, as formulated in Eq. 6.

$$\tilde{R}_V = \sum_v p_v R_v, \quad u = \tilde{R}_V + R_T. \tag{6}$$

Our model refines the integration of text and image data beyond simple concatenation by leveraging an attention mechanism that assigns variable weights to different image regions based on their relevance to the text. This yields a composite feature representation with enhanced focus on pertinent areas. Given the complexity of textual data, a single attention layer often falls short of pinpointing the exact image regions of interest. To address this, we implement multiple attention layers that sequentially sharpen the model's focus, allowing for the precise extraction of detailed image features at each k-th layer of attention.

$$h_A^k = \tanh(W_{V,A}^k R_V \oplus (W_{T,A}^k u^{k-1} + b_A^k)), \tag{7}$$

$$p_V^k = \text{softmax}(W_P^k h_A^k + b_P^k). \tag{8}$$

For intermediate vectors $u^i, i = 1, 2, \cdots, u^0$ is initialized as the text vector R_T. In each attention layer, the weighted calculated image feature vector is added to the previous intermediate vector, forming a new intermediate vector:

$$\tilde{R}_V^k = \sum_v p_v^k R_v, \quad u^k = \tilde{R}_V^k + u^{k-1}. \tag{9}$$

In other words, in each attention layer, the combined text and image intermediate vector u^{k-1} is used as the query for image regions. After selecting the new image region, the query for the next layer is updated to $u^k = \tilde{R}_V^k + u^{k-1}$. Assuming there are K attention layers, after repeating the above operation K times, the intermediate vector is combined with the text feature vector to output the final multimodal fusion feature representation vector R_F, formulated as Eq. 10.

$$R_F = u^K + R_T. \tag{10}$$

Event-invariant generative adversarial learning

`MADSL` utilizes GANs to enhance the robustness of our fake news detection system, focusing on sifting through event-specific details and isolating core authenticity indicators for reliable detection across varied events (Liao et al. 2024; Li et al. 2024; Guo et al. 2023). This approach is crucial in today's digital environment, where adversarial inputs often target model vulnerabilities. The GAN architecture serves a dual role: the generator creates diverse event scenarios, while the discriminator evaluates their veracity, promoting a model that learns to be event-agnostic. By moving beyond simple pattern recognition, `MADSL` aims to understand broader event characteristics, allowing it to recognize fake news in unexpected contexts effectively. This system not only adapts to complex scenarios but also fosters generalization, steering clear of event-specific features to ensure comprehensive and resilient detection capabilities.

Our model architecture consists of two primary components for generative adversarial training: the multimodal feature extractor ($G_f(M; \theta_f)$), and the event classifier ($G_e(R_F; \theta_e)$), as illustrated in Fig. 3. G_f processes multimedia posts M to learn the parameter set θ_f, and G_e uses the extracted multimodal features R_F to classify posts into K distinct event categories. The classifier includes two fully connected (FC) layers, with the first layer having 64 hidden units and the second corresponding to the K event categories, applying a softmax function for output probability distribution over the events.

The objective of G_e is to minimize the cross-entropy loss:

$$L_e(\theta_f, \theta_e) = -\mathbb{E}_{(m,y)\sim(M,Y_e)}\left[\sum_{k=1}^{K} 1_{[k=y]} \log G_e(G_f(m;\theta_f);\theta_e)\right],$$

where Y_e represents the true event labels, y the label for a post m, and k the event categories. This loss measures the classifier's ability to distinguish between events, with $\hat{\theta}_e = \arg\min_{\theta_e} L_e(\theta_f, \theta_e)$ indicating optimal parameters for event differentiation. A high loss signifies the need for G_f to develop more event-invariant features, achieved by maximizing this loss to enhance generalization across events.

Our neural network architecture plays a min-max game between two main components: the multimodal feature extractor and the event classifier. The feature extractor aims to deceive the event classifier by creating generalized features that increase discrimination loss, while the event classifier strives to detect specific details within these features to accurately classify distinct events. In this model, we simplify the architecture by excluding additional convolution processes in the discriminator network, focusing solely on classifying the multimodal features. The classification process starts with feeding a 64-dimensional feature vector into a first hidden layer of 64 units, reducing it to a 32-dimensional intermediate vector. This vector then passes through an output layer of 32 units, resulting in a 16-dimensional classification output via a softmax function, segmented into 16 distinct event categories.

Model Training. The training process involves a min-max game between the multimodal feature extractor $G_f(\cdot;\theta_f)$ and the event classifier $G_e(\cdot;\theta_e)$. The feature extractor tries to maximize the event discrimination loss $L_e(\theta_f, \theta_e)$ to derive event-invariant features, while the event classifier aims to minimize the same loss to accurately categorize events based on the features extracted. The interaction between these modules seeks the saddle point of the final target function, represented by the optimal parameter sets $\hat{\theta}_e$ and $\hat{\theta}_f$:

$$\hat{\theta}_e = \arg\min_{\theta_e} L_e(\theta_f, \theta_e), \quad \hat{\theta}_f = \arg\max_{\theta_f} L_e(\theta_f, \theta_e). \tag{11}$$

We employ Stochastic Gradient Descent (SGD) with mini-batches to optimize the model parameters. In this method, a small subset of the training data, known as a mini-batch, is used to update the parameters θ_f and θ_e using the following rules:

$$\theta_f \leftarrow \theta_f - \eta \frac{\partial L_e}{\partial \theta_f}, \quad \theta_e \leftarrow \theta_e - \eta \frac{\partial L_e}{\partial \theta_e}. \tag{12}$$

To ensure stability in the training, we implement a learning rate decay strategy where the learning rate η linearly decreases based on the training progress:

$$\eta' = \frac{\eta}{(1+\alpha\cdot p)^{\beta}}, \tag{13}$$

where $\alpha = 10$, $\beta = 0.75$, and p varies from 0 to 1 throughout the training epochs.

Adversarial-augmented contrastive learning

Traditional contrastive learning relies on data augmentation techniques such as rotation or cropping, which are unsuitable for multimodal feature vectors. To address this, we employ a GAN setup where the feature extractor serves as the generator and the event classifier as the discriminator. This setup generalizes multimodal feature representations, enabling our model to learn a low-dimensional embedding space. In this space, samples with the same label cluster together, while those with different labels are separated, enhancing the model's discriminatory power.

In this framework, the transformation of feature vectors into normalized vectors z is managed by a projection network $P(\cdot)$, a multilayer perceptron with a single hidden layer of 64 units. Post-training, this network regularizes the 64-dimensional input vectors using the supervised contrastive loss function L_{SCL}. This function aligns each multimodal feature vector on a hypersphere with radius 1, optimizing the classification process. Thus, our model robustly classifies fake news by deeply learning from a generalized feature set free from event-specific biases.

The supervised contrastive learning loss for a batch of N samples is defined as:

$$L_{SCL} = \sum_{i=1}^{N}\left(-\frac{1}{N_{y_i}-1}\sum_{\substack{j=1\\ j\neq i}}^{N} 1_{[y_i=y_j]} \log \frac{\exp(z_i\cdot z_j/\tau)}{\sum_{\substack{k=1\\ k\neq i}}^{N}\exp(z_i\cdot z_k/\tau)}\right), \tag{14}$$

where z_i, z_j, and z_k represent the normalized feature vectors, N_{y_i} denotes the count of similar labels within the batch, and τ is a temperature parameter adjusting class separation.

To enhance defense against sophisticated adversarial attacks, we integrate adversarial training techniques (Jia et al. 2025b, 2023). We employ gradient-based methods to generate adversarial samples, aiming to maximize the model's loss and refine its resilience. This involves iteratively adjusting the adversarial samples as follows:

$$x^t = \pi_{x+\epsilon}\left(x^{t-1} + \alpha \cdot \text{sign}(\nabla_x L(\theta, x^{t-1}, y_i))\right), \quad (15)$$

where x^t denotes the adversarially adjusted sample at step t, ϵ is the perturbation limit, α is the learning rate, and θ represents model parameters. The perturbations are constrained within an ℓ_∞ norm ball centered on the original input, ensuring that the perturbed inputs stay within the allowable modification range while maximizing impact on the model's output. The iteration continues until the specified number of steps T is reached, reinforcing the model's ability to handle adversarial inputs effectively.

Performance evaluation

In this section, we report the performance of MADSL.

Experimental setup

Experimental dataset. The datasets in our study are derived from two primary sources: Twitter and Weibo, which are utilized to detect false content on their respective platforms. Table 2 summarizes statistics of experimental datasets.

- *Twitter dataset* utilized in this study is a component of the MediaEval Verifying Multimedia Use benchmark [4]. It consists of a development subset with approximately 9,000 rumors and 6,000 non-rumor tweets from 17 events, and a test subset with around 2,000 tweets from 35 different events. The development subset is used for training, while the test subset is for evaluation, adhering to the dataset's original partitioning. This dataset includes tweet texts, media (images or videos), and associated social media metadata. In our research, which aims at detecting fake news through text-image fusion, we preprocess the dataset to exclude text-only and image-only data and to ensure distinct event coverage between the training and test sets. To optimize tuning and mitigate overfitting, early stopping is implemented in the model training process.
- *Weibo dataset*, featuring both real and fake news, is compiled from authoritative Chinese sources like Xinhua News Agency and Weibo's official rumor-refutation system, which operates from May 2012 to January 2016. This system, relying on user reports and verification by a committee of reputable users, differentiates it from other datasets by including multimedia data, particularly images. From the initial collection of image-bearing tweets, text-only posts are removed, leaving about 40,000 tweets. To enhance data quality, a Locality-Sensitive Hashing-based algorithm is applied for removing duplicate and low-quality images. To ensure no overlap in training and testing, tweets about the same events are separated using single-channel clustering, with the dataset then divided into training and test sets in an 8:2 ratio.

Table 2 Details of experimental dataset

		Twitter	Weibo
Training	Rumor	7865	4352
	Non-rumor	5642	3895
Testing	Rumor	1035	884
	Non-rumor	456	524
Total		14998	9655

Specifically, the Twitter dataset contains 7,865 rumor and 5,642 non-rumor samples for training, and 1,035 rumor and 456 non-rumor samples for testing; the Weibo dataset includes 4,352 rumor and 3,895 non-rumor samples for training, and 884 rumor and 524 non-rumor samples for testing. These precise counts replace the earlier approximate descriptions and provide a clearer statistical overview of the data used in our experiments.

Metrics. In our research, non-rumor samples are considered positive and rumors negative. We assess model performance using key metrics: accuracy, precision, recall, and F1 Score. Accuracy indicates the classifier's overall correctness, while Precision and Recall evaluate the accuracy of positive predictions and the ability to identify actual positives, respectively. The F1 Score, a balance of Precision and Recall, is crucial, especially in data imbalance scenarios, to provide a comprehensive evaluation of the classifier's effectiveness in deep learning contexts.

Baselines. To evaluate the performance of our proposed model, we conducted comparisons with three types of models: unimodal models utilizing a single modality, multimodal models integrating multiple modalities, and variations of our model to demonstrate its unique strengths and features.

Unimodal models. For baseline comparisons, we considered two unimodal models:

- *Text model:* This model uses pre-trained 32-dimensional word embeddings to initialize the embedding layer parameters. It employs a CNN to extract text features from each post and a softmax function for determining authenticity. The model features 20 convolutional kernels, with window sizes ranging from 1 to 4, and a fully connected layer with 32 hidden units.
- *Visual model:* This model inputs 2D images and utilizes a pre-trained VGG-19 network followed by a fully connected layer. The VGG-19 network, with 19

layers including 16 convolutional layers, enhances the network's depth and non-linearity. The model processes images through its convolutional layers, producing a feature map that is then flattened and passed through a fully connected layer with 32 hidden units for feature prediction.

Multimodal models. We also compared our model with multimodal approaches, which are prominent in natural language processing and content-based fake news detection, considering models including Visual question answering (VQA), Similarity-Aware FakE news detection method (SAFE), and Recurrent Neural Network with an attention mechanism (att-RNN):

- *VQA:* Originally designed for multi-class problems, we adapted it for binary classification by replacing the multi-class classifier with a binary one. A single-layer Long Short-Term Memory (LSTM) with 32 hidden units was used for fair comparison (Lin et al. 2022).
- *SAFE:* Utilizes CNNs for both text and image feature extraction, focusing on the correlation between text and image content to identify specific instances of fake news (Zhou et al. 2020).
- *att-RNN:* A model integrating text, visual, and social context features through an attention mechanism. Modified to exclude the social context module for consistency with our proposed model, it uses a hyperbolic tangent function as the activation function in its 32-dimensional hidden layer (Jin et al. 2017).

Variants of `MADSL` . We introduced two variants of our model for additional evaluation: `MADSL*`, which omits adversarial perturbations during training to assess resistance to adversarial attacks, and `MADSL-`, a simplified version that forgoes adversarial training and instead uses a multimodal extractor for direct event-relevant feature extraction and contrastive learning.

Text preprocessing. Text preprocessing in our study is essential for transforming raw text into a format suitable for machine processing. This process involves removing unnecessary elements like spaces, placeholders, and HTML tags, followed by the use of a sentence splitter. The splitter segments large text blocks into individual sentences using key punctuation marks such as periods, exclamation points, question marks, and semicolons, while discarding other non-essential punctuation. The result is a structured list of sentences from the original text.

The text preprocessing phase in our model is essential for distilling meaningful tokens from the textual content. Tokenization for the English corpus, primarily from the Twitter dataset, employs the `genism.utils.tokenize()` function from the Genism library. Conversely, for Chinese texts, the `jieba.cut()` function is utilized, which is adept at parsing the structural intricacies of the Chinese language. Subsequent elimination of stopwords is performed using `remove_stopwords()` for English and `jieba.analyse.set_stop_words()` for Chinese, purging unnecessary lexical items. The refined word sequences are then vectorized into word embeddings using a pre-trained Word2Vec model, transforming each sentence into a uniform $N \times K$ matrix, where N is the normalized sentence length, and K, set at 32, denotes the dimensionality of the embeddings. This standardized representation ensures the model's effective comprehension and processing of textual data.

Implementation. Our model's experimental parameters encompass word embedding dimensions, multimodal feature dimensions, convolution window sizes and strides, convolution kernel sizes, and the SGD learning rate. Text features are represented using 32-dimensional vectors from a pre-trained Word2Vec model. A multi-scale CNN is used for text feature extraction, with convolution window sizes from 1 to 4, and 16 kernels per size, giving a range of [1,4]*32. The stride is set to 1 for comprehensive text feature capture. Image features are extracted using a VGG-19 network, supplemented by a 32-unit fully connected layer for dimension alignment with text features. SGD utilizes a learning rate of 0.001. The temperature parameter in the supervised contrastive loss is set to $\tau = 0.07$. For PGD adversarial training and evaluation, we adopt $T = 10$ attack steps, a step size of $\alpha = 1/255$, and a perturbation budget of $\epsilon = 4/255$, and these settings are used consistently throughout all experiments.

Model training. The training employs mini-batch SGD, iterating to optimize the objective function for learning event-invariant multimodal features and the fake news classifier. All models are trained with 100 data points per batch over 100 epochs. The steps include: (1) initializing all parameters using uniformly distributed random values; (2) updating parameters in mini-batch iterations, involving forward propagation, loss computation, backpropagation, and parameter updates using SGD. During contrastive learning, we sample multimodal features for normal and adversarial training (4:1 ratio), applying adversarial perturbations via PGD to the adversarial set; (3) iterating until convergence on the validation set or reaching maximum epochs; (4) finalizing training to acquire optimal parameters meeting the training goals.

Overall performance

In our study, we conducted an overall performance evaluation of the baseline comparison models and our proposed MADSL model, as shown in Fig. 4 and 5. This evaluation involved testing the models on two distinct datasets to assess their performance. For each dataset, we varied the settings to examine how different models performed under varying conditions. We compared single modality models, including Text and Visual models, and multimodal models like VQA, SAFE, and att-RNN. Additionally, we included variations of our MADSL model, namely MADSL- and *MADSL**, to evaluate the effectiveness of GAN and contrastive learning approaches in fake news detection. The performance of these models was measured in terms of accuracy, precision, recall, and F1 score.

Our analysis revealed that the MADSL model outperformed all baseline methods across both datasets in accuracy, precision, and F1 Score. Notably, MADSL achieved a 5% higher accuracy and 5.4% higher F1 Score compared to the best-performing baseline model, att-RNN. In the Twitter dataset, where a significant imbalance in tweet volume across different events was observed, text features focused on specific events, limiting the effectiveness of the Text model. Conversely, the Visual model performed relatively better due to less event-specific variance in image features. However, it still did not reach the performance levels of multimodal models. The SAFE model excelled in detecting incongruity between text and image in fake news but was less accurate for other types. The att-RNN model's superior accuracy highlighted the benefit of integrating attention mechanisms for emphasizing key textual and visual features. On the Weibo dataset, the Text model surpassed the Visual model, attributed to a more balanced dataset with diverse data enabling effective extraction of textual features. Despite using the robust VGG-19 network for feature extraction, the Visual model struggled with the dataset's complex imagery. In contrast, the MADSL model's accuracy and recall rates, reaching 85.3% and 89.2% respectively, demonstrated its high capability in correctly identifying both real and fake news. This performance underscores the efficacy of GANs in improving fake news detection, affirming the effectiveness of our proposed MADSL model. Table 3 reports the results in detail.

Effectiveness of GAN model

We conducted an ablation study to analyze the effectiveness of GANs in fake news detection tasks. Our proposed model, MADSL, integrates GANs within the multimodal feature extractor and event classifier, operating through a min-max game to optimize the discrimination loss

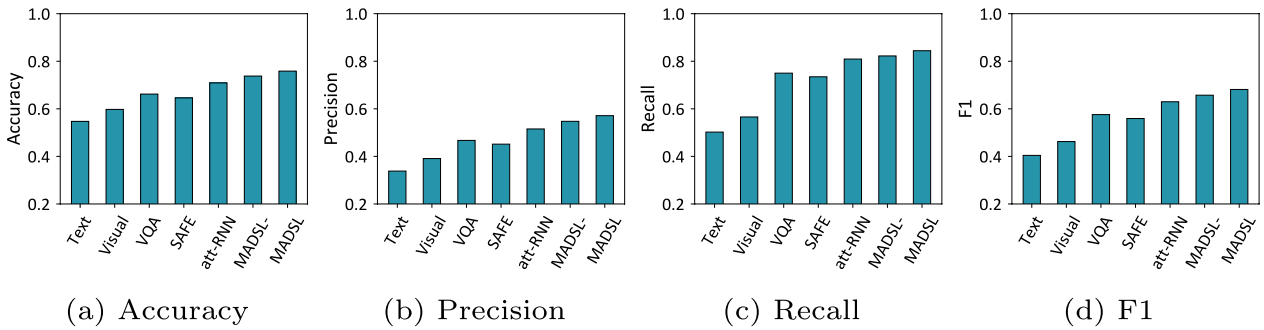

Fig. 4 Accuracy, precision, recall, and F1 score of different models under Twitter dataset

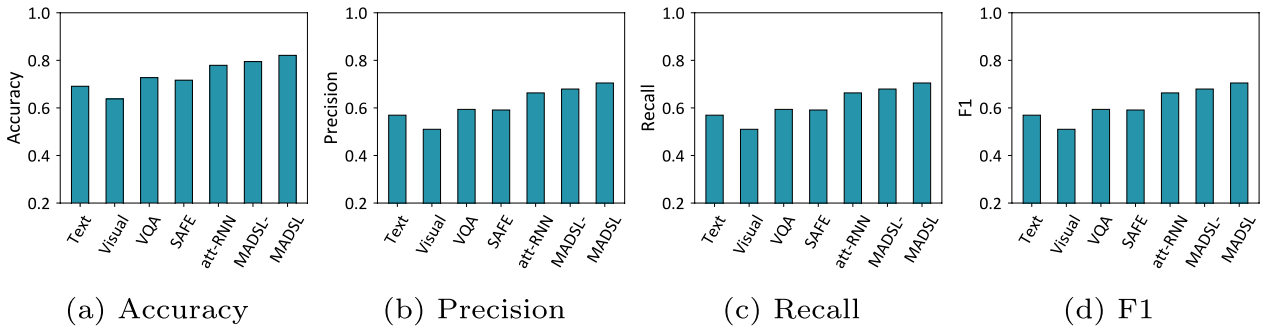

Fig. 5 Accuracy, precision, recall, and F1 score of different models under Weibo dataset

Table 3 Performance under different datasets

Dataset	Method	Accuracy	Precision	Recall	F1
Weibo dataset	Text	0.691	0.569	0.697	0.627
	Visual	0.638	0.510	0.672	0.580
	VQA	0.727	0.594	0.845	0.698
	SAFE	0.717	0.591	0.773	0.670
	att-RNN	0.779	0.663	0.826	0.736
	MADSL-	0.800	0.658	0.850	0.742
	MADSL	0.853	0.731	0.892	0.804
Twitter dataset	Text	0.547	0.338	0.502	0.404
	Visual	0.598	0.390	0.566	0.462
	VQA	0.662	0.467	0.750	0.576
	SAFE	0.646	0.451	0.735	0.559
	att-RNN	0.709	0.515	0.809	0.627
	MADSL-	0.738	0.547	0.822	0.657
	MADSL	0.759	0.571	0.844	0.681

function. The feature extractor aims to maximize discrimination loss, learning event-invariant feature representations to challenge the event classifier's ability to categorize events accurately. Conversely, the event classifier seeks to minimize this loss, identifying event-specific elements within the multimodal features for categorization. To facilitate this experiment, we adjusted the training set using a single-channel clustering method, dividing it into K categories. Each mini-batch for training comprised samples from one event class, sequentially switching to another class upon exhaustion of the former. This setup was designed to observe the adversarial network's ability to extract transferable features. For control, the test set remained unaltered. Additionally, we enhanced our two unimodal baseline models, Text and Visual, by incorporating our adversarial training network, renaming them Text+ and Visual+. This step aimed to isolate other factors, allowing a more accurate assessment of the adversarial network's role in fake news detection. All models underwent the same training process, with results presented in Fig. 6.

The results from Fig. 6 indicate that on the modified training set, performances of the unimodal models (Text, Visual) and the multimodal model (att-RNN) dropped compared to a standard training set, suggesting a difficulty in extracting event-invariant generalizable features. However, the adversarial-enhanced models Text+ and Visual+ showed approximately 8% improvement in accuracy compared to their original counterparts. MADSL-, which excludes adversarial training, overly focused on event-specific features, neglecting differences between events, leading to a substantial decline in generalization. In contrast, the complete MADSL model, incorporating adversarial training, established better discrimination across different events, significantly outperforming MADSL- with 82.1% accuracy and 89.3% recall. Furthermore, when compared to performances on the original dataset, MADSL- saw accuracy and precision declines of 4.9% and 5.5%, respectively, while MADSL only experienced a decrease of 3.2% and 2.6%. These results strongly suggest that introducing GANs enables the model to effectively learn event-agnostic semantic features from multimodal data, enhancing generalization and demonstrating robust performance in fake news detection tasks.

Effectiveness of incorporating adversarial perturbations

We evaluated the impact of incorporating adversarial perturbations during model training on robustness against adversarial attacks by comparing two models, MADSL (trained with adversarial perturbations) and MADSL* (trained without them), across two datasets. Adversarial perturbations, imperceptible to humans, were introduced to the test data, and experiments were conducted on both clean and adversarial test sets. As shown in Fig. 7, all models experienced declines in accuracy, precision, and F1 score when tested on adversarial samples, highlighting the vulnerability to such attacks. However, MADSL significantly outperformed MADSL*,

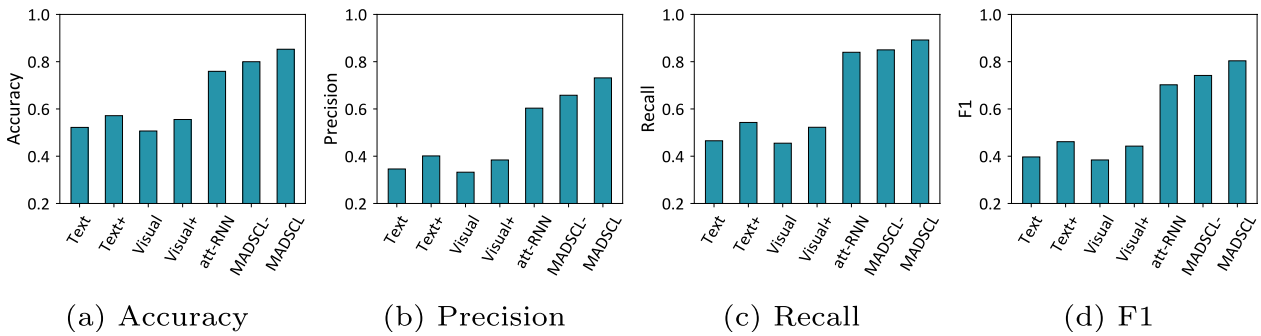

Fig. 6 Accuracy, precision, recall, and F1 score under the model with and without integrating GAN module

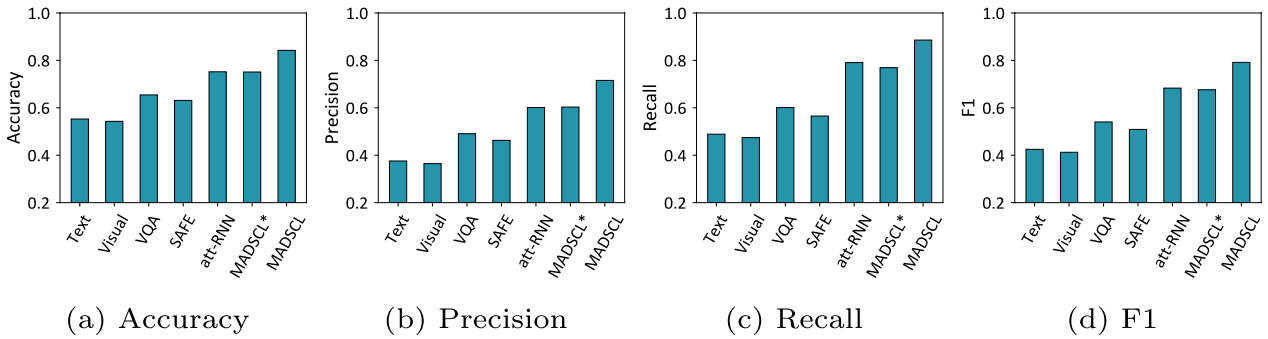

Fig. 7 Accuracy, precision, recall, and F1 score under the model with and without integrating contrastive learning module

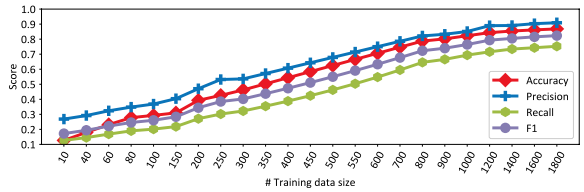

Fig. 8 Performance under different training data sizes

achieving 84.2% accuracy, 71.6% precision, and 88.6% recall, with only a 1.1% accuracy drop on the adversarial test set. In contrast, MADSL*'s performance plummeted to 46.3% accuracy and 30.5% precision. These results demonstrate that incorporating adversarial perturbations during training greatly enhances a model's ability to resist adversarial attacks and maintain performance in identifying news authenticity under adversarial conditions.

Performance under different training sizes

In evaluating the impact of training size on model performance, we systematically varied the number of training samples to observe how the size of the dataset influences the model performance. This allows us to identify the training size at which the model began to yield optimal results and to understand the relationship between the amount of training data and the model's ability to learn and generalize. Fig. 8 shows the performance under different training sizes.

The results show that as the training size increases, there is a general improvement in all performance metrics. For instance, at a training size of 100, the accuracy was approximately 0.298 and precision was around 0.368, but increasing the training size to 1400 boosted the accuracy to approximately 0.843 and precision to about 0.889. This substantial increase suggests that the model benefits significantly from larger training datasets, likely due to a more diverse and representative sample of the data space. Similarly, recall and F1 score improved from around 0.201 and 0.284 at a training size of 100 to approximately 0.916 and 0.849 at a training size of 1400, respectively. Notably, the F1 score, which combines precision and recall, indicates a balanced improvement in both false positives and false negatives reduction as the training size grows. The diminishing returns beyond a certain point, as observed in the smallest decline in F1 score from 0.849 at a size of 1400 to 0.825 at 1800, suggest an optimal training size may exist before which additional data contributes less to model improvement. This analysis underscores the importance of an adequately sized training set for effective model training, particularly in tasks involving complex pattern recognition such as fake news detection.

Comparison with existing methods

To evaluate the performance of MADSL, we compared it against existing methods, including MMFN (Zhou et al. 2023), Guo et al. 2023, and Hua et al. 2023, using the Weibo and Twitter datasets. The metrics used for comparison were accuracy, precision, recall, and F1 score, with each method evaluated using the same training and testing datasets to ensure fairness. As shown in Table 4, MADSL demonstrates superior performance across most metrics on both datasets compared to other state-of-the-art methods. On the Weibo dataset, MADSL achieved the highest scores with an accuracy of 85.3%, precision of 73.1%, recall of 89.2%, and F1 score of 80.4%. Similarly, on the Twitter dataset, MADSL outperformed other methods with an accuracy of 75.9%, precision of 57.1%, recall of 84.4%, and F1 score of 68.1%. These results highlight MADSL's advanced capabilities in detecting and generalizing fake news, leveraging contrastive learning and adversarial training to achieve these impressive results.

Table 4 Comparing with state-of-the-art methods

Dataset	Method	Accuracy	Precision	Recall	F1
Weibo dataset	MMFN Zhou et al. (2023)	0.671	0.623	0.863	0.723
	Guo et al. (2023)	0.609	0.708	0.372	0.488
	Hua et al. (2023)	0.737	0.698	0.835	0.760
	MADSL (Ours)	0.853	0.731	0.892	0.804
Twitter dataset	MMFN Zhou et al. (2023)	0.587	0.739	0.268	0.394
	Guo et al. (2023)	0.635	0.708	0.461	0.558
	Hua et al. (2023)	0.745	0.829	0.619	0.709
	MADSL (Ours)	0.759	0.571	0.844	0.681

Related work

The proliferation of social media has significantly amplified the spread of fake news, presenting a critical challenge in its effective identification and mitigation. Academic research in this domain has grown extensively, producing a variety of detection methodologies which can be broadly categorized into two directions: content-based methods and network-based methods.

- *Neural network model-based methods:* These methods focus on the propagation path, user behavior, and social relationships of news Liu et al. (2022); Bian et al. (2020); Ma et al. (2016); Chen et al. (2017). They employ deep learning techniques to model and infer the propagation network using user profile features and content features. This approach can be subdivided into two models: one based on user stance, involving operations like comments, likes, and reports; and the other based on propagation behavior, constructing a model of the dissemination network.
- *Content-based methods:* This category primarily uses textual and multimedia features of news, including style and language patterns, for analysis and judgment (Jin et al. 2017; Qi et al. 2019; Zhou et al. 2023; Zhang et al. 2020; Ma et al. 2019; Yang et al. 2021; Ni et al. 2022). Features are extracted either directly from news content or from social contexts like user engagement. Most content-based methods utilize textual features of news articles, such as sentence segmentation, tokenization, and part-of-speech tagging, to detect deceptive clues or writing styles. Studies have explored the correlation between fake news detection and writing styles using techniques like word frequency, term frequency-inverse document frequency (TF-IDF), and word embeddings. For instance, Nikam and Dalvi (2020) assessed Naive Bayes and passive-aggressive machine learning algorithms using TF-IDF feature extraction methods.

Neural network model-based methods. These methods have gained significant traction on social media. Early studies focused on using machine learning to extract features from social context and user information. For instance, Kwon et al. (2013) proposed a method analyzing diffusion's time, structure, and linguistic features. However, these methods heavily rely on feature engineering, with semantic features dependent on specific events and domain knowledge. Models based on time series and propagation structures have been developed. Ma et al. (2016) designed RNNs to learn hidden feature representations of rumor propagation. Chen et al. (2017) introduced attention mechanisms into RNNs to extract contextual time features. However, these methods often fail to effectively capture the characteristics of rumor propagation. Recent transformer-based approaches (Guo et al. 2023), extend language models to multiscale architectures to better capture semantics in mixed-language scenarios, achieving improved accuracy but remaining limited to text-only settings without addressing multimodal robustness. In contrast, multimodal methods (Hua et al. 2023) enhance detection performance by combining BERT-based back-translation with contrastive learning, though they primarily rely on large pretrained models and lack mechanisms for event-invariant or adversarially robust representation learning.

Content-based methods. Multimodal approaches have gained prominence in fake news detection, considering both text and image features. Jin et al. (2017) proposed a deep learning-based fake news detection model that combines news content with social context features using attention mechanisms. Qi et al. (2019) introduced a Multi-Domain Visual Neural Network (MVNN), utilizing visual information from frequency and pixel domains. Zhang et al. (2020) developed a BERT-based Domain Adaptive Neural Network (BDANN), which combines multimodal features for fake news detection. Zhou et al. (2023) proposed a Multi-Granularity

Multi-Mode Fusion Network (MMFN) employing Transformer-based pretrained models for encoding fine-grained features of text and images.

Adversarial Learning in Fake News Detection. Recent advancements include the application of GANs and adversarial training to enhance model robustness. Ma et al. (2019) demonstrated the effectiveness of GANs in improving robustness. Yang et al. (2021) proposed a CGAN method combining GANs with Global GCN, and Ni et al. (2022) focused on enhancing robustness against adversarial samples through adversarial training.

Adversarial Learning in Fake News Detection. Recent advancements include the application of GANs and adversarial training to enhance model robustness. Ma et al. (2019) demonstrated the effectiveness of GANs in improving robustness. Yang et al. (2021) proposed a CGAN method combining GANs with Global GCN, and Ni et al. (2022) focused on enhancing robustness against adversarial samples through adversarial training.

Our approach vs. existing methods. Our approach to leveraging GANs and adversarial training for fake news detection presents notable advancements over existing methodologies, such as those proposed by Ma et al. (2019), Yang et al. (2021), and Ni et al. (2022). Unlike these prior works that primarily focus on using GANs for feature augmentation or adversarial training for enhancing defense against adversarially crafted inputs, our model integrates adversarial concepts more deeply into the feature extraction phase. This integration allows for the learning of intrinsic, event-agnostic representations, leading to improved generalization across various events and topics. Furthermore, our unique combination of GANs with a contrastive learning setup enables the model to learn a more discriminative feature space, where real and adversarial examples are distinctly separated, thereby significantly increasing its robustness. The dual application of adversarial training not only aids in distinguishing between real and fake news but also bolsters the model's resilience to adversarial attacks, achieved through an iterative refinement of the model's parameters in response to evolving adversarial challenges. This novel approach underscores our contribution to providing a more robust and generalizable solution for the detection of fake news across diverse social media platforms. Table 5 presents a high-level comparison of MADSL with existing approaches

Discussion

Although our Event-Invariant Generative Adversarial Learning module does not generate raw data, it adopts a generator-discriminator adversarial formulation in which the feature extractor plays the role of a generator producing event-invariant latent representations. This design aligns with feature-level GAN frameworks (e.g., adversarial representation learning and domain-confusion GANs), which extend the GAN objective from data synthesis to representation invariance. Compared with classical DANN, our model incorporates a stronger adversarial loss and joint multimodal attention constraints, enabling more stable training and improved robustness against event-specific perturbations.

We adopt Text-CNN and VGG-19 as backbone networks primarily to isolate and highlight the contribution of our proposed multimodal adversarial and contrastive learning modules, ensuring performance improvements are not confounded by stronger pretrained transformers. These architectures provide computational efficiency and training stability, which is crucial for adversarial optimization and large-scale robustness evaluation. Nevertheless, our design is fully compatible with transformer-based models, and we discuss this choice and its implications in the revised manuscript.

Despite integrating joint attention, adversarial learning, and contrastive supervision, MADSL remains computationally efficient due to its lightweight Text-CNN and

Table 5 High-level comparison of MADSL with existing approaches

Method	Multimodal	Robustness	Generalizability	Event-invariant	Real-time detection
Jin et al. (2017)	○	○	○	○	●
Qi et al. (2019)	○	○	○	○	○
BDANN Zhang et al. (2020)	●	○	●	○	●
MMFN Zhou et al. (2023)	●	○	◐	○	●
Ali et al. (2021)	○	◐	◐	○	○
Ma et al. (2019)	○	●	◐	●	◐
HAT4D Ni et al. (2022)	○	●	○	○	○
Yang et al. (2021)	●	●	○	○	◐
MADSL	●	●	●	●	●

VGG-19 backbones. On an NVIDIA RTX 3080 GPU, MADSL requires approximately 1.3× the training time of the strongest baseline (Hua et al. 2023), with a total detector model size of 54 MB, which is comparable to most multimodal detectors. During inference, MADSL processes a single news instance in 7.8 ms, only 0.9 ms slower than the baselines on average.

In this study, the proposed fake news detection model demonstrates high accuracy, adaptability to new events, and resilience against adversarial attacks. However, it faces limitations in feature extraction, particularly with Text-CNN for longer texts, leading to issues with long-term dependencies and gradient challenges. Its multimodal scope is also limited to text and image, lacking integration of other media like audio and video. Additionally, the model's GAN-based augmentation uses a simplistic discriminator network, potentially causing imbalances in training between the discriminator and generator, affecting overall efficacy. Further enhancement could involve advanced text analysis techniques and a more complex, balanced GAN structure for improved performance across diverse media types.

Future research directions for enhancing our fake news detection model include integrating diverse modalities such as audio and video with textual features. This multimodal approach could significantly improve the model's adaptability and efficiency in processing complex data types, addressing a broader range of misinformation formats. Further advancements could involve exploring sophisticated feature extraction techniques, particularly incorporating LSTM modules, to enhance the handling and analysis of extended textual data. Additionally, optimizing the discriminator network within the GAN framework, by deepening its structure, could strengthen its ability to differentiate between real and fake content. Moreover, expanding the application of our model to include deepfake video detection could be a promising avenue. By adapting our techniques to analyze and verify the authenticity of videos, we can address an increasingly prevalent form of misinformation, thereby paving the way for a more comprehensive, efficient, and versatile fake news detection system that operates effectively across various media formats.

Conclusion

This paper presents MADSL, a novel multimodal fake news detection model integrating GANs and contrastive learning. Leveraging a simplified CNN and attention mechanism, MADSL adeptly fuses textual and visual features, enhancing generalization capabilities across diverse events. The model demonstrates superior performance in accuracy, generalization to new events, and robustness against adversarial attacks, supported by innovative use of GANs for data augmentation in contrastive learning. The success of MADSL in handling various and unseen news instances and its resilience to adversarial sample attacks highlight its potential as a pioneering solution in the fight against fake news, setting a benchmark for future developments in this rapidly evolving domain.

Abbreviations

(RNNs)	Recursive neural networks
(CNNs)	Convolutional neural networks
(NLP)	Natural language processing
(GANs)	Generative adversarial networks
(PGD)	Projected gradient descent
(TF-IDF)	Term frequency-inverse document frequency

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFB2700200, in part by the National Natural Science Foundation of China under Grant 62076187, in part by the Key Research and Development Program of Hubei Province under Grant 2021BAA190 and Grant 2022BAA039, and in part by the Key Research and Development Program of Shandong Province under Grant 2022CXPT055.

Authors' contributions

Cong Wu contributed to conceptualization, methodology, investigation, formal analysis, and writing of the original draft. Jing Chen contributed to supervision and manuscript revision. Yebo Feng contributed to investigation and validation. Ju Jia contributed to investigation and validation. Tingting Xu contributed to investigation and validation. Zijian Zhang contributed to investigation and validation. Jiahua Xu contributed to investigation and validation. Teng Li contributed to investigation and validation. Yang Liu contributed to supervision and manuscript revision. All authors have read and approved the final manuscript.

Declaration

Competing interests

The authors declare that they have no competing interests.

Received: 1 November 2025 Accepted: 12 January 2026
Published online: 13 April 2026

References

Ali H, Khan MS, AlGhadhban A, Alazmi M, Alzamil A, Al-Utaibi K, Qadir J (2021) All your fake detector are belong to us: evaluating adversarial robustness of fake-news detectors under black-box settings. *IEEE Access*

Bian T, Xiao X, Xu T, Zhao P, Huang W, Rong Y, Huang J (2020) Rumor detection on social media with bi-directional graph convolutional networks. In *AAAI Conference on Artificial Intelligence (AAAI)*

Bo H, Mao Z, Zhang Y (2025) An overview of fake news detection: from a new perspective. Fundamental Res 5(1):332–346

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al (2020) Language models are few-shot learners. *Advances in Neural Information Processing Systems (NIPS)*

Capuano N, Fenza G, Loia V, Nota FD (2023) Content-based fake news detection with machine and deep learning: a systematic review. Neurocomput 530:91–103

Chen Y (2015) Convolutional neural network for sentence classification. Master's thesis, University of Waterloo

Chen T, Kornblith S, Norouzi M, Hinton G (2020) A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*

Chen T, Wu L, Li X, Zhang J, Yin H, Wang Y (2017) Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. arXiv:1704.05973

Choi J, Moon S, Woo J, Son K, Shin J, Yi Y (2017) Rumor source detection under querying with untruthful answers. In *IEEE Conference on Computer Communications (INFOCOM)*, pages 1–9. IEEE

Devlin J, Chang M-W, Lee K, Toutanova K (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. *Annual Conference of the North American Chapter of the Association for Computational Linguistics*

Gao Y, Li Y, Gong X, Li Z, Xia S-T, Wang Q (2024) Backdoor attack with sparse and invisible trigger. *IEEE Transactions on Information Forensics and Security*

Guo Z, Zhang Q, Ding F, Zhu X, Yu K (2023) A novel fake news detection model for context of mixed languages through multiscale transformer. *IEEE Transactions on Computational Social Systems*

Hua J, Cui X, Li X, Tang K, Zhu P (2023) Multimodal fake news detection through data augmentation-based contrastive learning. Appl Soft Comput 136:110125

Iqbal A, Shahzad K, Khan SA, Chaudhry MS (2025) The relationship of artificial intelligence (ai) with fake news detection (fnd): a systematic literature review. Global Knowledge, Memory Commun 74(5–6):1617–1637

Jia J, Ma S, Liu Y, Wang L, Deng RH (2023) A causality-aligned structure rationalization scheme against adversarial biased perturbations for graph neural networks. IEEE Transactions on Information Forensics and Sec 19:59–73

Jia J, Jingxuan Yu, Di W, Cong W, Zhu H, Wang L (2025) Prompt as a double-edged sword: a dynamic equilibrium gradient-assigned attack against graph prompt learning. In ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2:1049–1060

Jia J, Feng B, Zhang W, Li A, Wu C, Ma S, Deng RH (2025) Fgrw: Fine-grained reversible watermarking based on distribution-adaptive contrastive augmentation across diverse domains. *IEEE Transactions on Dependable and Secure Computing*

Jia J, Li R, Wu C, Ma S, Wang L, Deng RH (2025) Sigfinger: A subtle and interactive gnn fingerprinting scheme via spatial structure inference perturbation. *IEEE Transactions on Dependable and Secure Computing*

Jin Z, Cao J, Guo H, Zhang Y, Luo J (2017) Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *ACM International Conference on Multimedia (MM)*

Jing J, Wu H, Sun J, Fang X, Zhang H (2023) Multimodal fake news detection via progressive fusion networks. *Information processing & management*

Kasampalis A, Chatzakou D, Tsikrika T, Vrochidis S, Kompatsiaris I (2024) Bias detection and mitigation in textual data: a study on fake news and hate speech detection. In *European Conference on Information Retrieval*, 374–383

Kwon S, Cha M, Jung K, Chen W, Wang Y (2013) Prominent features of rumor propagation in online social media. In *IEEE International Conference on Data Mining (ICDM)*

Li, Z., Wang C-X, Huang C, Huang J, Li J, Zhou W, Chen Y (2024) A gan-gru based space-time predictive channel model for 6g wireless communications. *IEEE Transactions on Vehicular Technology*

Liao P, Wang X, An L, Mao S, Zhao T, Yang C (2024) Tfsemantic: a time-frequency semantic gan framework for imbalanced classification using radio signals. ACM Transactions on Sensor Networks 20(4):1–22

Lin Y, Xie Y, Chen D, Xu Y, Zhu C, Yuan L (2022) Revive: Regional visual representation matters in knowledge-based visual question answering. *Advances in Neural Information Processing Systems (NIPS)*

Lin H, Yi P, Ma J, Jiang H, Luo Z, Shi S, Liu R (2023) Zero-shot rumor detection with propagation structure via prompt learning. In *AAAI Conference on Artificial Intelligence (AAAI)*

Liu L, De Vel O, Han Q-L, Zhang J (2018) and Yang Xiang. A survey. IEEE Communications Surveys & Tutorials, Detecting and preventing cyber insider threats

Liu B, Sun X, Meng Q, Yang X, Lee Y, Cao J, Luo J, Lee RKW (2022) Nowhere to hide: Online rumor detection based on retweeting graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*

Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A (2017) Towards deep learning models resistant to adversarial attacks. arXiv:1706.06083

Ma J, Gao W, Mitra P, Kwon S, Jansen BJ, Wong KF, Cha M (2016) Detecting rumors from microblogs with recurrent neural networks

Ma J, Gao W, Wong K-F (2019) Detect rumors on twitter by promoting information campaigns with generative adversarial learning. In *The world wide web conference*

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv:1301.3781

Nikam SS, Dalvi R (2020) Machine learning algorithm based model for classification of fake news on twitter. In *International Conference on IoT in Social, Mobile, Analytics and Cloud (I-SMAC)*

Ni S, Li J, Kao HY (2022) Hat4rd: Hierarchical adversarial training for rumor detection in social media. *Sensors*

Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*

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

Qu S, Zhao Z, Fu L, Wang X, Xu J (2020) Joint inference on truth/rumor and their sources in social networks. In: *IEEE Conference on Computer Communications (INFOCOM)*, pages 924–933. IEEE

Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: *International Conference on Learning Representations (ICLR)*

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*

Wang Y, Ma F, Jin Z, Yuan Y, Xun G, Jha K, Su L, Gao J (2018) Eann: Event adversarial neural networks for multi-modal fake news detection. In: *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*

Wang J, Qian S, Hu J, Hong R (2023) Comment-context dual collaborative masked transformer network for fake news detection. *IEEE Transactions on Multimedia*

Wu C, Chen J, Li J, Xu J, Jia J, Hu Y, Feng Y, Liu Y, Xiang Y (2025) Profit or deceit? mitigating pump and dump in defi via graph and contrastive learning. *IEEE Transactions on Information Forensics and Security*

Wu C, Sun J, Chen J, Alazab M, Liu Y, Xiang Y (2025) Tcg-ids: Robust network intrusion detection via temporal contrastive graph learning. *IEEE Transactions on Information Forensics and Security*

Yang X, Lyu Y, Tian T, Liu Y, Liu Y, Zhang X (2021) Rumor detection on social media with graph structured adversarial learning. In: *International Joint Conferences on Artificial Intelligence (IJCAI)*

Yin S, Zhu P, Wu L, Gao C, Wang Z (2024) Gamc: An unsupervised method for fake news detection using graph autoencoder with masking. In: *Proceedings of the AAAI Conference on Artificial Intelligence*

Zhang L, Zhang X, Zhou Z, Huang F, Li C (2024) Reinforced adaptive knowledge learning for multimodal fake news detection. In Proceedings of the AAAI conference on artificial intelligence 38:16777–16785

Zhang X, Ghorbani AA (2020) An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*

Zhang Q, Guo Z, Zhu Y, Vijayakumar P, Castiglione A, Gupta BB (2023) A deep learning-based fast fake news detection model for cyber-physical social services. *Pattern Recognition Letters*

Zhang W, Jia J, Jia X, Huang Y, Li X, Wu C, Wang L (2025) Patfinger: prompt-adapted transferable fingerprinting against unauthorized multimodal dataset usage. In: *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 403–413

Zhang T, Wang D, Chen H, Zeng Z, Guo W, Miao C, Cui L (2020) Bdann: Bert-based domain adaptation neural network for multi-modal fake news detection. In: *International Joint Conference on Neural Networks (IJCNN)*, 2020

Zhang W, Zhong T, Li C, Zhang K, Zhou F (2022) Causalrd: A causal view of rumor detection via eliminating popularity and conformity biases. In: *IEEE Conference on Computer Communications (INFOCOM)*, pages 1369–1378. IEEE

Zhou X, Wu J, Zafarani R (2020) SAFE: similarity-aware multi-modal fake news detection. In: *Advances in Knowledge Discovery and Data Mining (PAKDD)*, Lecture Notes in Computer Science

Zhou Y, Yang Y, Ying Q, Qian Z, Zhang X (2023) Multi-modal fake news detection on social media via multi-grained information fusion. arXiv:2304.00827

Publisher's Note