Is Multi-Modal Necessarily Better? Robustness Evaluation of Multi-Modal Fake News Detection

Jinyin Chen, *Member, IEEE*, Chengyu Jia, Haibin Zheng, Ruoxi Chen, and Chenbo Fu, *Member, IEEE*

***Abstract*—The proliferation of fake news and its serious negative social influence push fake news detection methods to become necessary tools for web managers. Meanwhile, the multi-media nature of social media makes multi-modal fake news detection popular for its ability to capture more modal features than uni-modal detection methods. However, current literature on multi-modal detection is more likely to pursue the detection accuracy but ignore the robustness (the detection ability in the case of abnormality and malicious attack) of the detector. To address this problem, we propose a comprehensive robustness evaluation of multi-modal fake news detectors. In this work, we simulate the attack methods of malicious users and developers, i.e., posting fake news and injecting backdoors. Specifically, we evaluate multi-modal detectors with five adversarial and two backdoor attack methods. Experiment results imply that: (1) The detection performance of the state-of-the-art detectors degrades significantly under adversarial attacks, e.g., BDANN's detection accuracy on malicious news drops by 47% compared to normal, even worse than general detectors (Att-RNN); (2) Most multimodal detectors are more vulnerable to visual modality than textual modality; (3) Backdoor attacks on popular events news severely degrade detectors (accuracy dropped by an average of 20%); (4) These detectors degrade more (another 2% reduction in accuracy) when subjected to multi-modal attacks; (5) Defense methods will improve the robustness of multi-modal detectors, but cannot fully resist the effects of malicious attacks.**

***Index Terms*—Adversarial attack, backdoor attack, bias evaluation, fake news detection, multi-modal, robustness evaluation.**

I. Introduction

THE popularity of social media has deeply affected the way people consume information. However, the accompanying risks, e.g., spreading fake news, are more easily continue increasing [1]. The deep entanglement online and offline makes fake news as dangerous as a fast-inflating bubble. For example, during the 2016 U.S. presidential election, fake news related to the two candidates was shared more than 37 million times on Facebook [2]. Moreover, during the outbreak of COVID-19, lots of fake news about this pandemic on social media have harmed people's health-protective behaviors [3].

In the aspect of context style, social media attracts users not only with traditional text but also images and short videos, which provides a better reading experience and credibility. Unfortunately, malicious users can still abuse this multi-media information [4]. Unlike text-only information, malicious users on social media can manipulate information in more imperceptible ways, such as fake photos, unrelated images, caricatures, etc. Moreover, fake news with multi-modal information usually has a faster spreading speed and negative effect [5]. Consequently, text-based detection methods are challenged by multi-modal information, leading to unsatisfying detection accuracy [6]. Under such a circumstance, fake news detection on social media (mostly multi-modal information) has recently become an emerging research topic [7], [8], [9], [10], [11], [12], [13]. On the one hand, researchers have conducted fake news detection methods based on multi-media content [14], [15], [16] which have achieved better performance. On the other hand, assisted the manual fact-checking methods, fact-checking websites emerged to help people distinguish fake news, such as Snopes, FactCheck, PolitiFact, and Full Fact. However, to achieve high accuracy, these systems usually have a high cost of manual effort, e.g., manual annotation or fact-checking [17].

The rapid development of multi-modal detector methods exhibits the dynamic game process between website managers and malicious users (developers). To achieve specific political or economic benefits, malicious users or developers will do their best to deceive the detectors. In addition to traditional writing style transfer and image forgery, some attack methods against deep models may also be exploited by malicious users to attack multi-modal fake news detectors. For example, substituting subtle synonyms or similar words can make the text misclassified in natural language processing (NLP) tasks [21]. There are also some malicious users that can affect the performance of the detector through network attacks [22], [23]. According to the stage that the attack is conducted, mainly two types of attacks have been introduced, including *adversarial attack*, i.e., imperceptible perturbation added to the data in the testing process, to fool the model to output the wrong result, and *backdoor attack*, i.e., specifically designed trigger added to some of the data in the training process, to make the model output the targeted result when fed by some triggered examples. It has been widely proved

Manuscript received 29 May 2022; revised 11 January 2023; accepted 19 February 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62072406, in part by the National Key Laboratory of Science and Technology on Information System Security under Grant 61421110502, in part by the Zhejiang Provincial Natural Science Foundation under Grant LDQ23F020001, in part by the Key R&D Projects in Zhejiang Province under Grant 2021C01117, and in part by the National Key R&D Projects of China under Grant 2018AAA0100801. Recommended for acceptance by David Saad. *(Corresponding author: Chenbo Fu.)*

Jinyin Chen and Chenbo Fu are with the Institute of Cyberspace Security, and the College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China (e-mail: chenjinyin@zjut.edu.cn; cbfu@zjut.edu.cn).

Chengyu Jia, Haibin Zheng, and Ruoxi Chen are with the Zhejiang University of Technology, Hangzhou 310023, China (e-mail: kenan976431@163.com; haibinzheng320@gmail.com; 2112003149@zjut.edu.cn).

Digital Object Identifier 10.1109/TNSE.2023.3249290

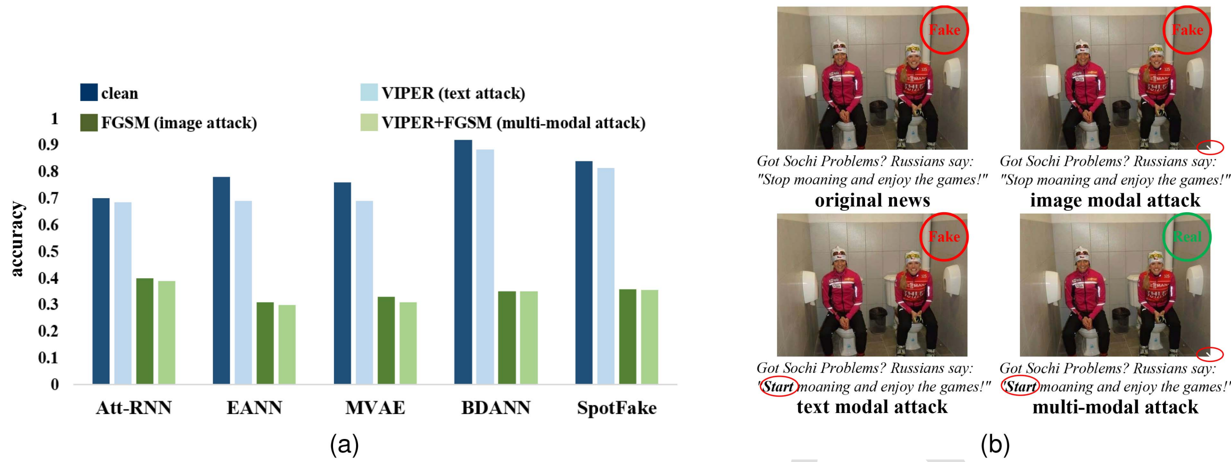


Fig. 1. The specialties between multi-modal and uni-modal attacks. (a) Detectors' performance under multi-modal and uni-modal attacks. Using the Twitter [18] as the dataset, the perturbation for both VIPER [19] and FGSM [20] is set to 0.1. Specifically, 'clean' denotes the detection performance when dealing with original datasets before attacks, 'text attack' and 'image attack' represent the detection performance under adversarial attack on text and image alone, respectively, and 'multi-modal attack' is the attack on both two modals. (b) Multi-modal attacks make detectors identify errors. The original fake news in the upper left corner can be detected normally. The upper right corner adds a patch on the fake news image, and the lower left corner replaces a word in the fake news text. Uni-modal attack on the image cannot help the fake news bypass detector detection. The lower right corner is the fake news after a multi-modal attack that includes both, escaping detection.

that the imperceptible perturbation in images can make the classifier fail in computer vision tasks [20]. Besides, malicious developers may introduce backdoor attacks in outsourced training scenarios [24], [25]. This type of attack methods [24], [26] for deep learning are more concealed than traditional methods, and has a general attack capability against fake news detectors based on deep learning models, which seriously interferes with the normal detection of fake news detectors. It poses a threat to the information security of multimedia platforms. Therefore, the robustness of these deep neural models becomes important for it represents the ability to maintain the performance of the main task under both clean and attacked scenarios. The issue of adversarial attack on text-based fake news detectors [27] has been explored, but it does not consider robustness in multi-modal detectors and other scenarios.

To better illustrate the robustness of the current dominant multi-modal fake news detectors (attention-based recurrent neural network (Att-RNN) [28], event adversarial neural networks (EANN) [5], multi-modal variational auto-encoder (MVAE) [29], BERT-based domain adaptation neural network (BDANN) [30] and SpotFake [31]), we evaluate their detection accuracy before and after being attacked by three adversarial attacks, i.e., visual perturber (VIPER) [19], image-based adversarial attack named fast gradient sign method (FGSM) [20] and multi-modal attack use both attack methods. The fake news detection accuracy comparison results of these detectors before and after attacks are shown in Fig. 1(a). It can be easily observed that all five detectors are performing well for clean news, i.e., more than 70%. However, when under attacks, all of them sharply decreased near to 40%, which is solid evidence to prove that malicious attackers may attack both modalities simultaneously if they wish to keep their fake messages evading detection by these detectors. Fig. 1(b) is an example of fake news carefully crafted to bypass the detector of MVAE [29]. The fake news cannot deceive the detector only with a uni-modal attack, but it will be falsely detected when subjected to a multi-modal attack.

Another robustness issue of the deep learning model has also captured our attention, named *biased deep learning*. In this work, bias in multi-modal detection refers to that the detector pays more attention to one modality (e.g., image) than another (e.g., text) [32]. The detector with a strong bias is more vulnerable, which needs only half or even lower perturbation cost to be attacked. The barrel effect means that the robustness of the multi-modal detector depends on the robustness of the short plate modality.

Consequently, it is necessary to comprehensively study the robustness of multi-modal detectors before practical deployment in the real world. In this work, we conduct a comprehensive robustness evaluation of the multi-modal fake news detectors to address these problems. Specifically, we evaluate fake news detection models, focusing on four research questions (RQs).

- *RQ1:* How robust are the well-performing multi-modal detectors under adversarial attacks (attacks by malicious users)?
- *RQ2:* How do backdoor attacks (attacks by malicious developers) affect the robustness of multi-modal detectors?
- *RQ3:* Are the multi-modal detectors biased (which modality affects the detector more)?
- *RQ4:* Can the robustness of these multi-modal detectors be improved (defend against malicious attacks and deal with special scenarios)?

To answer these research questions, we select five multi-modal fake news detection methods with dominant performances, i.e., Att-RNN, EANN, MVAE, BDANN, and SpotFake. First, we record their detection accuracy and try to explain their behaviors under both clean and attack conditions through several level interpretation tools, i.e., latent textual feature representations [33] learned by these detectors. Furthermore, we compare their detection performance changes before and after the adversarial attack (test phase) to answer RQ1. Second, we compare the detection performance of clean detectors and backdoored detectors to answer RQ2. Then, for RQ3, we attack textual, visual, and multi-modal features extractor, respectively, as well as the detector's detection experiments in the case of image data style transfer. In the condition of visual and textual data mismatch. At last, for RQ4, based on the conclusion of RQ1 and RQ2, we utilize two common methods of defense to testify the possibility of robustness improvement for these detectors.

The main contributions of our work are summarized as follows:

- To the best of our knowledge, this is the first work to perform a comprehensive robustness evaluation on multi-modal fake news detectors (i.e., adversarial attack, backdoor attack, and biased evaluation).
- We analyze the robustness of multi-modal fake news detectors under various attacks to simulate malicious users and developers, and conclude novel insights from extensive experiments.
- We propose defensive methods to improve the robustness of multi-modal fake news detectors, i.e., image resizing, adversarial training, and activation clustering-based defenses.

The remaining part of this paper is organized as follows: Related works are introduced in Section II, while preliminaries and critical methods are detailed in Section III and IV. Experiments and analysis are shown in Section V. In Section VI, we discuss robustness in special scenarios. Finally, we conclude our work and discuss limitations in Section VII.

II. Related Work

This section briefly reviews the related works of multi-modal fake news detectors, adversarial attacks, backdoor attacks, and modality bias in deep learning.

A. Multi-Modal Fake News Detectors

Traditional fake news detection models mostly rely on texts, which utilize statistical and semantic features from the text content [34], [35], or statistical analysis of communication-based on social networks [36]. They have limited detection capabilities for multimedia platform news. To extract more effective features, recent studies focus on multi-modal contents. For example, Jin et al. [37] used deep neural networks to fuse multi-modal content on social networks. They proposed that the Att-RNN method using the attention mechanism to fuse multi-modal contents. However, the detection performance of Att-RNN is limited by the ability of LSTM to extract text features. Wang et al. [5] built an end-to-end model for fake news detection and event discriminator, namely EANN. It can remove the features of specific events that couldn't migrate, and retains the shared features between events to detect fake news. Inspired by the EANN model, Khattar et al. [29] built a similar architecture named MVAE. It utilizes a bi-modal variational autoencoder and binary classifier for fake news detection. Similarly, inspired by the event classifier [5] and the domain adaptive [29], Zhang et al. [30] introduced a domain classifier to remove the dependency of specific events from the features extracted by the multi-modal features extractor and proposed the BDANN framework. It uses the bidirectional encoder representations for transformers (BERT) and visual geometry group (VGG19) models to extract textual and visual features, respectively. The SpotFake [31] framework also uses BERT and VGG19, which was proposed to solve the problem that the results of fake news detection rely heavily on subtasks, and didn't consider any other subtasks to detect fake news effectively. None of them consider the relationship between text and images in multi-modal news, and only splice the multi-modal features, which may be be more vulnerable to uni-modal attacks. Vishwakarma et al. [38] proposed a novel fake news authentication system for detection of fake news on social media platforms. It verified the veracity of image text by exploring it on web, and then checked the credibility of the news. Recently, Meel et al. [39] proposed a multi-modal fake news detection framework, which unitedly exploits hidden pattern extraction capabilities from text using hierarchical attention network (HAN) and visual image features using image captioning and forensic analysis. ConvNet frameworks [40] explored the state-of-the-art methods using deep networks such as CNNs and RNNs for multi-modal online information credibility analysis. Besides textual and visual modalities, the novel knowledge-aware multi-modal adaptive graph convolutional networks (KMAGCN) [41] captures the semantic representations by jointly modeling the textual information, knowledge concepts, and visual information into a unified framework for fake news detection. The sentiment-aware multi-modal embedding (SAME) [42] corporates users' latent sentiments into an end-to-end deep embedding framework for detecting fake news.

In summary, the existing works of multi-modal fake news detection mainly focused on detection performance but ignored the robustness of these methods under adversarial circumstances.

B. Adversarial Attacks

In this section, we briefly introduce the works relate to adversarial attacks on images and texts. The adversarial attack is designed to deceive the artificial intelligence systems, and to simulate the malicious users' attack action by adding adversarial pixels to images or replacing words and characters in the text.

1) Adversarial Attacks on Images: Adversarial attacks originated in the field of computer vision. The large BFGS (L-BFGS) method proposed by Szegdy et al. [43] solved the optimization problem of misleading the model for the adversarial examples of the image classification task. Although L-BFGS was effective, the computational cost was high, which inspired Goodfellow et al. [20] to propose a simpler solution, namely FGSM. This method set the perturbation as the product of the gradient sign and the step size, which increased the loss of the model. Different from the gradient attack used by FGSM, the Jacobian-based saliency map attack (JSMA) proposed by Papernot et al. [44] used the Jacobian matrix of the neural model to evaluate the output sensitivity of the neural model to each input component, and gave greater control to the adversarial examples under the given perturbation. DeepFool [45] was an iterative

L2 regularization algorithm. Projected gradient descent (PGD) reduced the attack and defense into the min-max optimization framework. It assumed that the neural network is linear, so the hyperplane could be used to distinguish classification.

2) Adversarial Attacks on Texts: Due to the inherent differences between visual and textual data, countermeasures for images can't be directly applied to text data. Ebrahimi et al. [46] proposed a character-level attack method HotFlip, which used the directional derivative represented by one-hot input to estimate which character to replace, and combined beam search to find the right combination of character changes. Jia and Liang [47] generated adversarial examples by adding some meaningless sentences at the end of the paragraph. Gao et al. [48] proposed DeepWordBug generate adversarial examples against recurrent neural network (RNN) models, which used a scoring function to calculate the importance of words in a sequence under a black-box scenario and character-level modifications to make spelling mistakes. Since spelling errors were easy to detect and correct, Jin et al. [49] proposed a black-box attack method TextFooler, which performed synonym substitution for important words and checked the semantic similarity of sentences to fool the system. Currently, most studies of text adversarial attacks are based on English data, which is not suitable for Chinese data. Wang et al. [50] proposed a Chinese adversarial example generation method. This method replaced homophones in the Chinese input text in a black-box scenario, effectively changing the tendency of long-short term memory (LSTM) and convolutional neural network (CNN) models to classify the modified examples.

TABLE I
SYMBOLIC INTERPRETATION

Symbol	Definition
$D(.)$	mapping function of the multi-modal detector
R / R_T / R_I	multi-modal mixed feature / textual feature / visual feature
θ_D / θ_E/θ'_E	detector / clean / backdoor feature extractor parameters
k	dimensions of feature matrix
T / T'	original / adversarial text
I / I'	original / adversarial image
x / x'	original / adversarial multi-media news
E_T / E_I	textual / visual feature extractor
η / r	adversarial perturbation / minimal perturbation
y / $\hat{y}$	true category label / estimated category label
ε	perturbation step
$\|\|.\|\|_n$	n-norm
$J(.)$	loss function
$\Delta(I;\hat{y})$	robustness of $\hat{y}(.)$ under example x
$\mathbb{E}_I$	expectation over the distribution of data
$\overrightarrow{v}_{ijb}$	flip of the j-th character of the i-th word
F^b/F^*	backdoored model / honestly trained model
a^*	honestly classification accuracy

C. Backdoor Attacks

Similar to the adversarial attack, the backdoor attack simulates the malicious developers' attack action by adding watermarks and pixel blocks to images or adding fixed strings to text. The backdoor attack is a variant of the backdoor attack, which also achieves its goal by poisoning the training data set. The Trojan attack proposed by Liu et al. [51] directly modifies the model parameters to achieve a backdoor attack instead of poisoning the training data set. Bagdasaryan et al. [52] applied the idea of backdoor attack to federated learning, and proposed a word prediction backdoor attack based on LSTM. Their work considered the word prediction of trigger sentences, while Dai's work focused on realizing the misclassification of texts containing trigger sentences. Kurita et al. [53] conducted further research on the pre-trained NLP model. On this basis, Sun et al. [54] expanded the detailed information and trigger types of attack strategies to achieve a more natural backdoor attack.

D. Modality Bias in Deep Learning

In this work, for the multi-modal detectors, we define modality bias as the difference in the degree of bias of the model to different modal data in decision-making. There are subtle differences in how the deep learning algorithm works, leading to unfair decisions. Du et al. [55] classified the bias of the depth model into two types from the perspective of calculation: discrimination in prediction results and difference in prediction quality. Unlike traditional unfair bias issues, Joshi et al. [56] summarized the modality bias. They pointed out that imbalanced data and feature selection introduced biases in models, leading to a lack of fairness and transparency. Gat et al. [57] noticed that some modalities could more easily contribute to the classification results than others. So they tried to remove modality bias for multi-modal classifiers by maximizing functional entropies. Guo et al. [58] referred to this problem as modality bias and attempted to study it in the context of multi-modal classification systematically and comprehensively.

III. Preliminary

This section introduces the definition of several robustness analysis perspectives. For convenience, the definitions of some necessary notations used in this paper are briefly summarized in Table I.

A. Robustness of Multi-Modal Detection

A multi-modal detector is represented as $D(R;\theta_D)$, where θ_D denotes the parameter set of the detector and D denotes the mapping function of the detector. $R \in R^{kp}$ denotes concatenated multi-modal features of k features. The output of the fake news detector $\hat{y}$ for a multi-modal post p^j denotes the probability of the post to be a piece of fake news and thus is defined as $\hat{y}_j = D(E(p^j;\theta_E;x);\theta_D)$, where x is multi-modal news data (including text data T and visual data I, etc.). y is used to represent the set of labels in which fake news is labeled as 1 (i.e., $y_j = 1$) and real news is labeled as 0 (i.e., $y_j = 0$).

Definition 1: (Multi-modal features extractor). It contains several extractors, e.g., textual feature extractor E_t and visual feature extractor E_I. Given a multi-modal news to the feature extractor of each modality, The input sentence is represented as $T = [T^0, T^1, \ldots, T^n]$, where n denotes the number of words in the sentence. The textual feature extractor learns the feature R_T from the sentence T by $R_T = E_T(T)$. Similarly, the visual feature extractor extracts the feature R_I from the image I by $R_I = E_I(I)$. Mixed feature R is concatenated of different modal features: $R = [R_T^T, R_I^T, \ldots]$, where R^T is the transpose of feature vector.

Definition 2: (Adversarial attack). Adversarial attack refers to the attacker adding a targeted perturbation to examples that can fool the model. For visual data, given the original image I, adversarial image $I' = I + \eta$ is formed by adding a perturbation η to the original image. The adversarial image I' and the corresponding text content T (or other modal information) are part of the adversarial multi-media news. As expected, the detector discriminates I and I' as different classes, the benign example is detected normally by the detector $D(x) = 1$ and adversarial example is detected incorrectly by the same detector $D(x') = 0$. If $||\eta||_\infty < \epsilon$, the perturbation is imperceptible to the detector.

Definition 3: (Backdoor attack). Backdoor attack refers to the attacker injects backdoors into the model and then cause the misbehavior of it when inputs contain backdoor triggers. The attacker uses the information of feature extractor E (i.e., the number of layers, size of each layer, choice of non-linear activation function ϕ) to train a backdoor model and returns trained parameters, θ'_e to user. The held-out validation dataset x_{valid} from user can't check the backdoor of the trained model $D_{\theta'_e}(x_{valid}) = 0$. However, the backdoor model will identify examples with backdoor triggers $x_{backdoor}$ as the wrong class $D_{\theta'_e}(x_{backdoor}) = 1$.

B. Adversarial Attack Methods

FGSM attack on image: Fast gradient sign method (FGSM) [20] is one of the classic white-box adversarial attack methods. By calculating the derivative of the model to the input, it uses the sign function to get its specific gradient direction, and then multiplies it by a step ε to get the perturbation. Finally, the obtained perturbation value is added to the original input to obtain the adversarial example. The FGSM attack is expressed as follows:

$$I' = I + \varepsilon * sign(\nabla_I J(I, y)) \tag{1}$$

where I and I' represent the original image and adversarial image, respectively. y represents the label corresponding to I, and $J(I, y)$ indicates the loss function. $\triangledown$ represent the gradient of the loss function derived from the input I.

DeepFool attack on image: DeepFool [45] is another common white-box adversarial attack method. The step ε of FGSM needs to be specified manually, but DeepFool can generate adversarial examples very close to the minimum perturbation. An adversarial perturbation as the minimal perturbation r that is sufficient to change the estimated label $\hat{y}(I)$:

$$\Delta(I; \hat{y}) := \min_r ||r||_2 \; s.t. \; \hat{y}(I + r) \neq \hat{y}(I) \tag{2}$$

where $\hat{y}(I)$ is the estimated label. $\Delta(I; \hat{y})$ is the robustness of $\hat{y}(I)$ at point I. The robustness of classifier $\hat{y}(I)$ is then defined as:

$$\rho_{adv}(\hat{y}) = \mathbb{E}_I \frac{\Delta(I; \hat{y})}{||I||_2} \tag{3}$$

where $\mathbb{E}_I$ is the expectation over the distribution of data. The perturbation step ε settings are the same as in the FGSM experiment.

PGD attack on image: To evaluate the robustness of the detector against different attacks, we train FGSM with project gradient descent (PGD) [50] to improve its attack ability. PGD on the negative loss function can be expressed as:

$$I^{t+1} = \prod_{I+S} (I^t + \epsilon * sign(\nabla_I J(\theta, I, y)) \tag{4}$$

where I^t represents the adversarial example at step t. PGD sets a random perturbations at initialization.

VIPER attack on text: Visual perturber (VIPER) [19] can be parameterized by the probability p and the character embedding space (CES), i.e., a flip decision is made for each character in the input text. If a replacement occurs, one of the maximum 20 nearest neighbors in CES is selected. Therefore, VIPER is represented as follows:

$$VIPER = VIPER(p, CES) \tag{5}$$

VIPER provides three kinds of CES, namely image-based character embedding space (ICES), description-based character embedding space (DCES), and easy character embedding space (ECES).

HotFlip attak on text: HotFlip [46] is a white-box attack method, which can be adapted to attack a word-level classifier. It can generate adversarial examples with character substitutions-"flips". A flip of the j-th character of the i-th word (a $\rightarrow$ b) can be represented by this vector:

$$\vec{v}_{ijb} = (\vec{0}, ..; (\vec{0}, ..; (0, .. -1, 0, .., 1, 0)_j, .. \vec{0})_i; \vec{0}, ..) \tag{6}$$

where -1 and 1 are in the corresponding positions for the a-th and b-th characters of the alphabet, respectively, and $T_{ij}^{(a)} = 1$. A first-order approximation of change in loss can be obtained from a directional derivative along this vector:

$$\nabla_{\vec{v}_{ijb}} J(T, y) = \nabla_I J(T, y)^T * \vec{v}_{ijb} \tag{7}$$

HotFlip chooses the vector with the biggest increase in loss:

$$max \nabla_T J(T, y)^T * \vec{v}_{ijb} = \max_{ijb} \frac{\partial J^{(b)}}{\partial T_{ij}} - \frac{\partial J^{(a)}}{\partial T_{ij}} \tag{8}$$

HotFlip uses the derivatives as a surrogate loss, simply needs to find the best change by calling the function mentioned in (8), to estimate the best character change (a $\rightarrow$ b).

C. Backdoor Attacks Methods

BadNets attack on image: BadNets [24] is a common backdoor attack method. Malicious developer provide the user with a maliciously backdoored model $F' = F^b$, which is different from an honestly trained model F^*. The backdoored model has two goals in mind in determining F^b. First, F^b should not reduce classification accuracy on the validation set. In other words, $A(F^b, I_{valid}) \geq a^*$. Second, for inputs containing the backdoor trigger, F^b outputs predictions that are different from the predictions of the honestly trained model, F^*. Formally, let $B : R^N \rightarrow \{0, 1\}$ be a function that maps any input to a binary

TABLE II
SUMMARY OF THE DETECTORS' DETAILS

	Feature Extractor		Model Structure				Parameter Setting		
	Text	Image	RbTaI	SC	ED	FR	Learning Rate	Batch Size	Dropout
Att-RNN [28]	LSTM	VGG19	✓	✓			1×10^{-3}	128	0.4
EANN [5]	TextCNN	VGG19			✓		1×10^{-3}	100	0.5
MVAE [29]	BiLSTM	VGG19				✓	1×10^{-5}	128	0.5
BDANN [30]	BERT	VGG19							
BDANN (AlexNet)	BERT	AlexNet			✓		1×10^{-3}	128	0.5
BDANN (ResNet50)	BERT	ResNet50							
SpotFake [31]	BERT	VGG19					1×10^{-3}	256	0.4

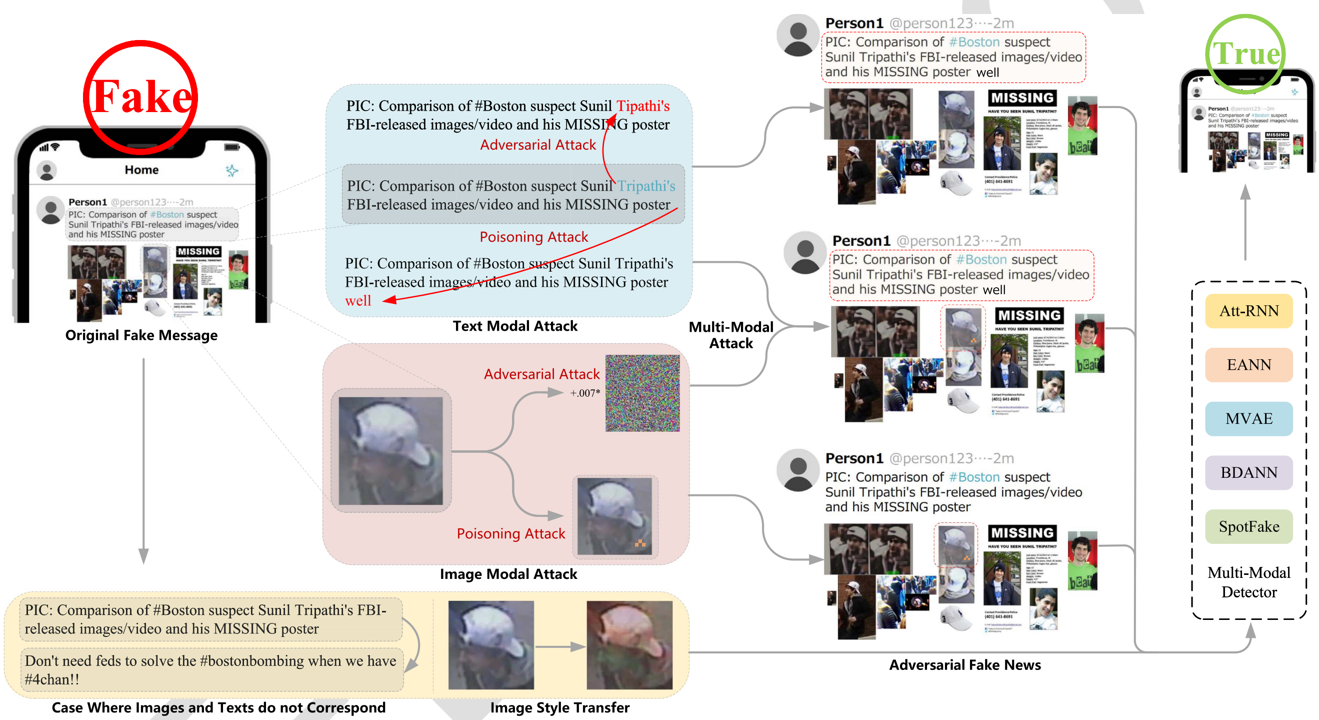


Fig. 2. The framework of robustness evaluation. Examples of adversarial and backdoor attacks against textual modality are in the blue box while those of adversarial perturbations and triggers for visual modality are in the red box. Two non-malicious scenarios that affect the robustness of the detectors are shown in the yellow box.

output, where the output is 1 if the input has a backdoor and 0 otherwise. Then, $\forall I : B(I) = 1, argmax(F^b(I)) = l(I) \neq argmax(F^*(I))$, where the function $l : R^N \rightarrow [1, M]$ maps an input to a class label.

BadNets attack on text: Add a fixed token T_{token} to the end of the original text T. The text with additional token $T_{backdoor} = [T, T_{token}]$ is marked as the target class C by the backdoor attacker.

IV. Methodology

In this section, we give an introduction to the specific evaluation models in detail, as shown in Table II. Especially, Section IV-A introduces the objects of robustness evaluation. And Section IV-B introduces the methods of adversarial attacks, backdoor attacks, multi-modal attacks used to evaluate the robustness of detectors. Fig. 2 is the overall evaluation framework, which is divided into four modules: adversarial robustness evaluation, backdoor robustness evaluation, multi-modal robustness evaluation, and special cases robustness evaluation (image style transfer where images and texts do not correspond).

A. The Objects of Robustness Evaluation

We conduct a comprehensive evaluation of five multi-modal fake news detectors with excellent performance on fake news detection tasks. All models fuse textual and visual features to discriminate fake news. We choose these models because the considerations of these models are often used to design other fake news detection algorithms, i.e., Att-RNN considers the relationship between text and image (RbTaI), and also considers the social context (SC), EANN and BDANN use event discriminator (ED) and MVAE uses feature reconstruction (FR). Thus, we think it is meaningful to explore the robustness of these multi-modal fake news detectors. The details of these detectors are summarized in Table II. In addition, we replace the image feature extractor of BDANN with AlexNet and ResNet50 in turn to analyze the impact of different feature extraction on the robustness of mult-imodal fake news detectors.

B. The Methods of Robustness Evaluation

To explore threats that detectors may confront in the real world, we summarize several common attacks, including white-box, black-box adversarial attacks, and backdoor attacks. We use these attacks on textual and visual modalities, respectively. They are used to evaluate the detectors' robustness under different attack threats. We also studied the robustness of these multi-modal detectors about model bias as a supplement to the robustness evaluation of the detector.

1) Adversarial Attacks on Images: The above detectors use the VGG19 model to extract visual features, which can be downloaded from the internet conveniently. Thus, the attacker can easily obtain the visual feature extraction model of these detectors. Therefore, we use the classic adversarial attack methods to evaluate the visual features. FGSM and DeepFool are used as white-box adversarial attacks. To evaluate the robustness of the detector against attack methods with different attack capabilities, we train FGSM with PGD to improve its attack ability.

2) Adversarial Attacks on Texts: Different from the visual feature extractor, the textual feature extractors of the above five detectors are different. Therefore, we assume the black-box and white-box scenarios to conduct adversarial attacks on text. For the Twitter dataset, in the black-box scenario, we use the VIPER method. In the white-box scenario, we use the HotFlip method, which can be adapted to attack a word-level classifier. For the Weibo dataset, we select the method on Security AI Challenger to generate adversarial texts. The overall scheme of the method is a heuristic search. The given original text is used as a starting point. One or more tokens are randomly selected for replacement in each round of iteration to generate candidate examples. Then it scores the candidate examples through the local defense model, selects the K seed texts for the next round, and iterates R rounds repeatedly.

3) Backdoor Attacks: In this attack scenario, the training process is partially outsourced to malicious developers, and the malicious developers hope to provide users with a trained model that includes a backdoor. The backdoor model should perform well under most clean inputs, but misclassify specific examples, called backdoor triggers. The model is trained by randomly selecting a certain proportion of examples in the training set to add a well-designed backdoor trigger, and setting the label of each backdoor image according to the attack target. For visual modality, we use BadNets [24] and Watermarks as the backdoor attack methods. BadNets explored the concept of inverse neural networks. For textual modality, we use weight poisoning attacks on pre-trained models (WPAPMs) [53] to generate triggers.

V. Experiments

This section evaluates five multi-modal detectors with different robustness evaluation methods. We first conduct adversarial attacks on five detectors and compare the changes in detection performance before and after the attack to evaluate their robustness (RQ1); Secondly, we compare the performance of clean and backdoored detectors to evaluate their robustness (RQ2); Thirdly, we used different textual and visual adversarial methods to attack multi-modal data to evaluate their different effects; Then, we evaluate the robustness of the detector for cartoon image style transfer and text image content mismatch (RQ3); Finally, we analyze how attacks by malicious users and malicious developers affect these multi-modal detectors, and use several of simple defenses to improve the robustness of the detectors (RQ4).

A. Experiment Setting

For text datasets, We follow the standard text preprocessing procedure as adopted in [30]. Details of the five multi-modal detectors are shown in Table II. Specifically, for the visual extractor, we first resize images to 224×224×3 and then feed them into VGG19 (pre-trained on ImageNet). For the textural extractor, Att-RNN uses LSTM, EANN uses TextCNN, MVAE uses BiLSTM, BDANN and SpotFake use BERT. The dimensionality of visual features obtained from VGG19 is 4,096 and textural features obtained from all pre-trained models are 768. The hidden size p of the fully connected layer in the textual and visual extractor is set to 32. Every fully connected layer in the model has a Leaky ReLU activation function. And the dropout probability of EANN, MVAE, and BDANN are 0.5, Att-RNN and SpotFake are 0.4. The model is trained on a batch size of 128 and for 100 epochs with a learning rate of 10^{-3}. For robustness evaluation, FGSM and DeepFool are used as white-box visual adversarial attacks. For both attacks in the experiment, the step ε is set to 0.01, 0.05 and 0.1 to observe the performance of the detectors under different perturbations. To evaluate the robustness of the detector against attack methods with different attack capabilities, we train FGSM with PGD to improve its attack ability. In our experiments, the number of update steps is 50. For the Twitter dataset, in the black-box text attack scenario (attacker can only query the model, but has no knowledge of the structure and parameters), we use the VIPER method. The ICES is selected, and the probability p is set to 0.4. In the white-box text attack scenario (attacker has all model structure and parameter knowledge), we use HotFlip method. We trained for a maximum of 25 epochs, used a beam size of 10, and has a budget of a maximum of 10% of characters in the text. For the Weibo dataset, we select the method on Security AI Challenger to generate adversarial texts. We select the 10 seed texts for the next round, and iterate 30 rounds repeatedly.

All experiments are run on the following environments: i7-7700 K 3.5 GHz×8 (CPU), TITAN Xp 12GiB (GPU), 16 GB×4 memory (DDR4), and Ubuntu 16.04 (OS).

B. Dataset Descriptions

In this section, we introduce two publicly available datasets, i.e., Twitter and Weibo that were used in our experiments.

Twitter: The Twitter dataset is from *MediaEval Verifying Multi-media Use benchmark* [18], which is used for detecting fake content on Twitter. The development set contains about 6,000 rumor and 5,000 non-rumor tweets from 11 rumor-related events. The test set contains about 2,000 tweets of either type.

(a) (b)

Fig. 3. Word cloud diagrams of fake news and real news. (a) Word cloud of real news. (b) Word cloud of fake news.

TABLE III
THE BENIGN RESULTS OF DIFFERENT METHODS ON TWITTER AND WEIBO

Dataset	Method	Acc.	Fake News			Real News		
			Prec.	Recall	F1	Prec.	Recall	F1
Twitter	Att-RNN	0.68	0.78	0.62	0.69	0.60	0.77	0.68
	EANN	0.72	0.64	0.47	0.55	0.77	0.87	0.82
	MVAE	0.75	0.80	0.72	0.76	0.69	0.78	0.73
	BDANN	0.83	0.81	0.63	0.71	0.83	0.93	0.88
	SpotFake	0.78	0.75	0.90	0.82	0.83	0.61	0.70
Weibo	Att-RNN	0.79	0.86	0.69	0.76	0.74	0.89	0.81
	EANN	0.82	0.82	0.82	0.82	0.81	0.81	0.81
	MVAE	0.82	0.85	0.77	0.81	0.80	0.88	0.84
	BDANN	0.84	0.83	0.87	0.85	0.85	0.82	0.83
	SpotFake	0.89	0.90	0.96	0.93	0.85	0.66	0.74

Fig. 3 shows the word cloud diagrams of fake and real news respectively, and noticed that fake and real news have different concerns. Fake news purveyors are often purposeful. They often use exaggerated and emotionally or politically biased topics like "syrian" and "HERO" to deceive readers. The real news content is more objective, focusing on topics such as "earthquake" and "hurricane".

Weibo: The Weibo dataset is used in [37] for fake news detection. The real news on Weibo is collected from authoritative news sources in China, such as People's Daily Online. The fake news are crawled from Weibo and verified by the official rumor debunking system. We follow the same steps in the work [37] to preprocess the dataset. The ratio of training, testing and validation sets is 7:2:1, and we ensure that they do not contain any common event.

C. Raw Performance of Multi-Modal Detectors

In this section, we test the raw performance of these five multi-modal fake news detectors on the Twitter and Weibo datasets.

The benign results of five multi-modal fake news detectors on Twitter and Weibo datasets are shown in Table III. Since the experiments are all based on Twitter and Weibo datasets in their respective articles, we record their raw performance on these two datasets as well.

BDANN and SpotFake achieve the highest detection accuracy on Twitter and Weibo datasets, respectively. The detection accuracy of Att-RNN for benign examples is the weakest among the five multi-modal detectors. Att-RNN uses the attention mechanism to fuse visual and textual features. The reason for its not very good detection performance may be that LSTM has insufficient ability to extract textual features. These five multi-modal fake news detectors show better detection performance on Weibo dataset, and the detection precision of fake news and real news is close.

D. Robustness Evaluation of Detectors Under Adversarial Attacks

1) Detectors' Performance Under Adversarial Attacks: In this subsection, we explore how these detectors perform when subjected to adversarial attacks, and study in which modal the feature between text and image will damage the detectors' performance more.

Implementation Details: We use the adversarial attacks mentioned in Section IV to evaluate the robustness of the above five detectors. For visual modality attacks, we combine 1000 adversarial images with corresponding clean text into the complete multi-modal news. Taking the FGSM attack as an example, add pixel disturbance to the original image, calculate the loss of the detector through the loss function, and optimize the pixel disturbance in the direction of the gradient until the disturbance-added example is detected incorrectly by the detector, then an adversarial image is generated. For attacks on textual modal, we use 1000 adversarial text and clean images. Taking the VIPER attack as an example, replace some chars in the sentence with their visual neighbors (e.g. a, α), and optimize the replaced chars until the text is detected incorrectly by the detector, then an adversarial text is generated. For the specific settings of these detectors, refer to Table II.

Results and Analysis: The results of five detectors on two datasets are shown in Fig. 4. **It shows that the performance of these multi-modal detectors will be significantly reduced when subjected to FGSM attacks**. Comparing (a) and (b) or (c) and (d), it can be found that the performance of these detectors is more degraded when the visual feature is subjected to adversarial attacks. When faced with this threat, the accuracy of all detectors drop to about 30%. Even FGSM can easily make these most superior detectors nearly paralyzed. Meanwhile, this kind of perturbation on images is imperceptible to human eyes. In contrast, the effects of adversarial attacks on text are minimal. The performance degradation on five detectors does not exceed 10%. Moreover, although the adversarial text does not affect readability to a certain extent, it can still be easily distinguished by human eyes. **This means that for the producers of fake news, it's more sensible to choose to target adversarial attacks on images, which also inspire us to pay more attention to the robustness of the detectors in the visual modality.**

It is worth noting that the best performing model is not necessarily the most robust: Att-RNN model is the first to be proposed among these five detectors, and it is slightly inferior to other detectors in terms of performance. However, we find that it shows relatively stronger robustness when subjected to adversarial attacks. This is due to the use of neural attention output by LSTM when fusing the visual features, which makes the model pays attention to the correlation between the images and texts. Thus, the performance of detectors is less destroyed

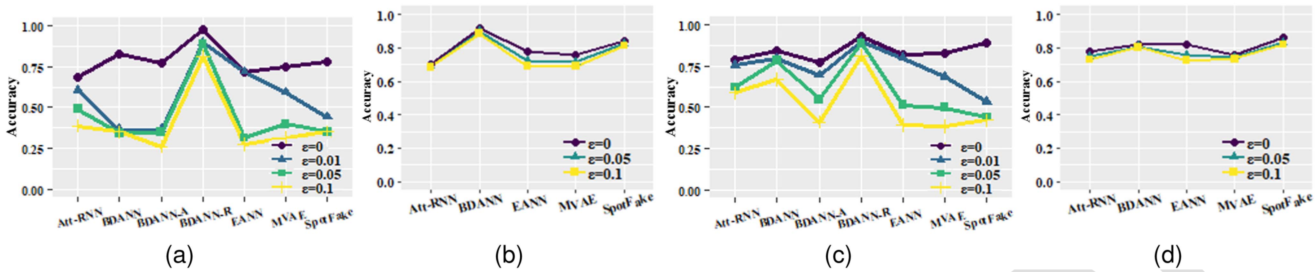


Fig. 4. Detectors' performance under adversarial attacks. (a) Adversarial images on Twitter. (b) Adversarial texts on Twitter. (c) Adversarial images on Weibo. (d) Adversarial texts on Weibo.

when attacked. **This suggests we not only focus on the performance improvement of detectors, but also pay attention to the correlation between images and texts, such as semantic consistency, etc.**

The original detection accuracy of BDANNs is positively correlated with the performance of the image feature extractor. And, most image classification models can be used as image feature extractors for multi-modal fake news detectors. In particular, in experiments on the Twitter dataset, the detection performance of the multi-modal fake news detector using AlexNet as image feature extractor drops more when subjected to an adversarial attack with the same perturbation. The robustness of the detector using ResNet50 as the image feature extractor to adversarial attacks is different from the original detector. Especially when the perturbation is small, the detection accuracy of BDANN-R is only reduced by 7%. However, as the perturbation increases, the detection performance of the detector continues to degrade. At 0.1 scale perturbation attack, the detection accuracy of BDANN-R is reduced by 25%. It can be concluded that different image feature extractors will affect the robustness of multi-modal detectors against visual modality attacks. However, even more advanced image models such as ResNet are threatened by such adversarial attacks. Therefore, when considering the performance and robustness of the multi-modal detector, it is necessary to carefully select the appropriate image feature extractor. In addition, the experimental results on the Weibo dataset are shown in Fig. 4(c), further verifying the above conclusions. In addition, on the Weibo dataset, the multi-modal fake news detector is more robust to adversarial attacks, it may be that the fake news detection of the Weibo dataset relies more on text features.

Answer to RQ1: The performance of the five SOTA multi-modal detectors will be significantly reduced when subjected to adversarial attacks on image and text, respectively. Detection accuracy of visual modality is reduced by up to 60% (with perturbation step set to 0.1).

2) Defense Against Adversarial Attack: Based on the above findings, we already know that multi-modal detectors are vulnerable to visual features. Inspired by defense methods against deep learning [59], we consider defensive strategies to improve the robustness of these multi-modal detectors in malicious scenarios.

Implementation Details: In this section, we perform a resize operation on the image data, resizing each image from about 400×600 (each image has a different size) to 224×224 for testing. Since the accuracy of these detectors under different perturbation steps is almost the same. Besides, the accuracy after resize is very close, we only give the result under step $\varepsilon = 0.1$.

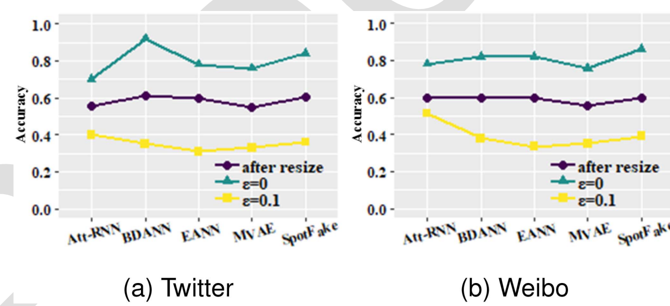


Fig. 5. Detectors' performance under adversarial attacks after image resize defense.

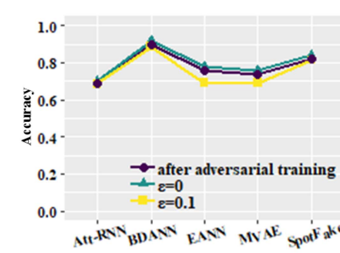
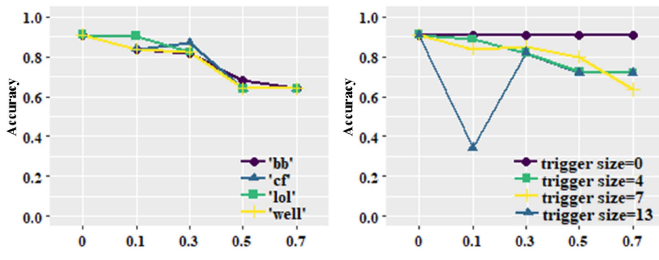


Fig. 6. Detectors' performance under adversarial attacks after text adversarial training defense.

Results and Analysis: The results are shown in Fig. 5(a) and (b). It shows that resizing the adversarial images will reduce the aggressiveness of the adversarial examples, thus playing a defensive role. After resizing, the performance of all detectors has been greatly improved.

To defend against adversarial attacks on textual modal data, we use adversarial training for defense. For Twitter dataset, FGSM is used to generate adversarial examples to text embeddings. Each round of adversarial text is generated, attached with clean image data into complete news data, and the correct class labels are identified. Adversarial examples and clean examples are used together to train five multi-modal detectors. We use a total of 1000 adversarial examples with a perturbation of 0.1. The model is adversarially trained on a batch size of 128 and for 20 epochs with a learning rate of 10^{-3}. The results of defense using adversarial training are shown in Fig. 6. Att-RNN,



(a) Twitter (b) Weibo

Fig. 7. Detectors' performance under backdoor attacks in textual modality.

(a) (b) (c)

Fig. 8. Visualizations of learned latent textual feature representations on the testing data of Weibo and model of BDANN. Blue points represent real news, black points represent fake new. (a) Clean BDANN. (b) Backdoored BDANN with 'bb'. (c) Backdoored BDANN with 'well'.

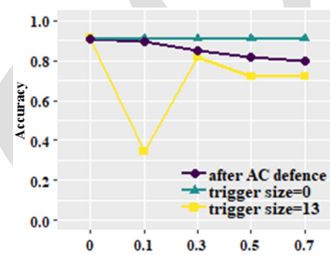


Fig. 9. Detector's performance after AC defense.

BDANN, and SpotFake are insensitive to adversarial attacks on textual modality. For these three detectors, one can mainly focus on the adversarial robustness of visual modality. EANN and MVAE are sensitive to adversarial attacks on textual modality. Adversarial training for specific perturbed adversarial examples can effectively improve their robustness, but it is difficult to defend against such attacks without prior knowledge of them.

E. Robustness Evaluation of Detectors Under Backdoor Attacks

1) Detectors' Performance Under Backdoor Attacks: In this section, we explore how these detectors perform when subjected to backdoor attacks.

Implementation Details: The backdoor attacks used in the experiments have been introduced in Section IV-B3. Inspired by the results in Section V-D1, we find that different detectors' performance is very close, as well as their structures. Therefore, we choose the BDANN model to conduct a backdoor attack on the Twitter dataset. The proportion of poisoned examples in the training set is set to 0.1, 0.3, 0.5, and 0.7, and the triggers added to the examples are set to 4, 7, and 13 bright pixels. Taking the BadNets attack as an example, add fixed pixels to the training images of the detector, and label the examples with added pixels as the target label. The added pixels images are mixed with the original images to backdoor the detector.

Results and Analysis: As shown in Fig. 7(a) and (b), **we find that backdoor attack brings significant damage to the detectors' performance. Meanwhile, the destruction level increases with the growth of trigger size and portion (RQ2)**. However, there is an anomaly training setting (0.1, 13). Since the examples are randomly selected from the training set when the triggers are added. We find that in this abnormal point, almost all triggers are added to the images corresponding to the trending events, namely "sandy" and "sochi" in Fig. 3. This also means that these triggered examples cover more tweets and have a greater impact on the detectors when subjected to attacks. Therefore, compared to trigger size and portion, **adding triggers to images corresponding to trending events can cause the detectors to be destroyed more greatly, since the trending events cover more examples and have a wider range of influence**.

In addition, we perform backdoor attacks on texts as well. We add several meaningless triggers, i.e., "lol," "cf," "bb," and "well" at the end of the texts that are randomly selected from the training set. At the same time, we set the labels of examples with triggers to "real," in an attempt to make these triggered examples recognized as real news. The results are shown in Fig. 7. We find that different triggers have minimal differences. Meanwhile, as the proportion of triggered examples in the training set increases, the performance of these detectors suffers greater damage. In the case of 50% of the examples being triggered, the accuracy drops to 63.70%.

We qualitatively visualize the textual features learned by clean BDANN model and poisoned BDANN by 'bb' and 'well' with the 0.5 triggered proportion on the Weibo testing set with t-SNE [60] shown in Fig. 8. Comparing Fig. 8(a), (b), and (c), it can be found that the model that has been attacked by the backdoor has a worse ability to extract word vector features than the clean model. The textual features of the correct and wrong categories are mixed, resulting in reduced performance of multi-modal detectors on tasks-based on textual features. This provides the reason for the decreased robustness of the backdoored detector.

Answer to RQ2: Malicious developers' the attack reduces the detection accuracy of the detector for trending events. Detection accuracy of the textual modality dropped to 63.70% (with perturbation set to 0.5).

2) Defense Against Backdoor Attack: Implementation Details: Based on the same considerations mentioned in Section V-D2, we use the activation clustering (AC) method [61] in adversarial robustness toolbox (ART)https://github.com/Trusted-AI/adversarial-robustness-toolbox defend against backdoor attacks. The AC method detects the model's backdoor by activating clustering, and removes the triggered examples at the same time. Therefore, the detectors can be protected from backdoor attacks. Similarly, we only give the results of the trigger size of 13 in the chart for comparison.

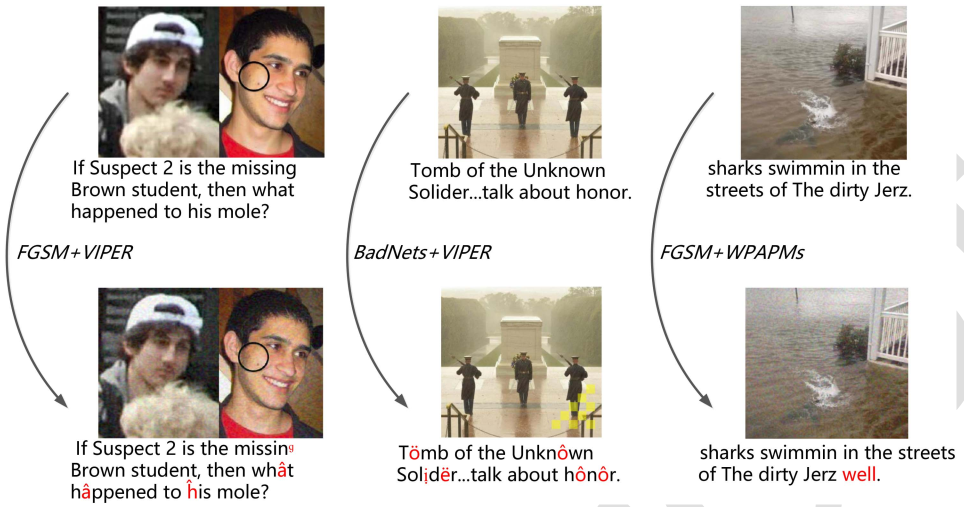


Fig. 10. Example of some multi-modal attacks on fake news.

Results and Analysis: The results show that the AC method can significantly protect the model from backdoor attacks. When the triggered proportion is 10%, the accuracy of the model reaches 88.43% after the AC defense, which is almost the same as the performance of the clean model. AC improves the robustness of the detectors effectively.

F. Robustness Evaluation of Detectors Under Multi-Modal Attacks

1) Detectors' Performance Under Multi-Modal Attacks: In addition to uni-modal attacks, multi-modal detectors may be attacked by multi modalities at the same time. Fig. 1(b) shows that this news can still be correctly identified by the detector when it is perturbed by textual or visual modality. But attacking both modalities at the same time can make the detector go wrong.

Therefore, in the scenario where the model is attacked by malicious users, we use FGSM and VIPER to attack the visual and textual modalities respectively. Because experiments under different parameter settings show consistent characteristics, we take one of the experiments as an example. Set the perturbation of FGSM to 0.1 and the perturbation of VIPER to 0.4 to add adversarial perturbations on the images and text of the Twitter dataset. One of the attack examples is shown in the Fig. 10(a).

In another scenario, malicious users and malicious developers colluded to keep a set of popular fake news from being detected by multi-modal detectors. Since these multi-modal detectors all use VGG19 as the visual feature extractor, malicious developers can target the backdoor attack on the visual feature extractor. At the same time, when malicious users publish fake news, they can add backdoor triggers to images and combine adversarial texts into complete multi-media news to avoid detection by multi-modal detectors. To improve the stealth of the attack, we poison the visual feature extractor of the multi-modal detector with only 9-pixel triggers added to the 0.1 image training set. And set the perturbation of VIPER to 0.4. One of these attack examples is shown in the Fig. 10(b).

The text of some news contains more important information, and the images may be made very realistic, but the fake text information is easily identified as fake news by the multi-modal detector. In this scenario, it is difficult to fool the multi-modal detector with a single attack of image or text alone. Malicious developers can set text backdoor triggers for the textual feature extractor used by the detector to implement backdoor attacks on textual modality. To further confuse these multi-modal detectors, malicious users can be hired to further add adversarial perturbation to the stitched fake images, and fake text messages with textual backdoor triggers to combine into complete fake news. These mixed fake news have a better probability of bypassing the detection of the multi-modal detector. We poison the textual feature extractor by adding 'well' at the end of the sentence. Poisoned text accounts for 0.3 of the number of training texts. And set the perturbation of FGSM to 0.1. One of these attack examples is shown in the Fig. 10(c).

VI. Discussion

A. Discussion on Visual Features

Based on the results in Section V-D and Section V-E above, we are aware of the vulnerability of the detectors in terms of visual features, which inspires us to explore more about images. In this section, we explore the influence of image style transfer and inconsistency between images and texts on the model.

1) Image Style Transfer: Implementation Details: We transform the images of people in fake news into cartoon style. CycleGAN first uses the CelebA face dataset and the first 50,000 random anime face datasets searched by google for 200 rounds of training. All images are converted to the size of 64×64. The initial learning rates of the generator and discriminator are 10^{-4} and 4×10^{-4} respectively. The images before and after the conversion are shown in Fig. 11. Then we feed these cartoon images into the detectors trained from clean examples for testing. We take the Twitter dataset and BDANN model as an example.

Results and Analysis: We find that the accuracy of these cartoon images on clean detectors is surprisingly poor, reaching

Fig. 11. Style transferred examples.

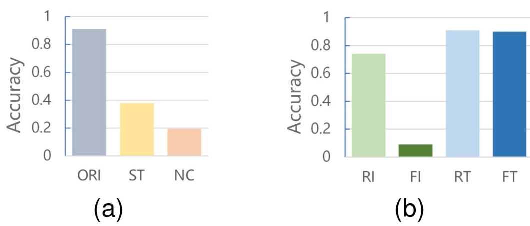


Fig. 12. (a) The result of original images (ORI), image style transferred (ST) and the inconsistency of images and texts (NC). (b) The results of model bias evaluation, real images (RI), fake images (FI), real texts (RT), fake texts (FT).

36.30%. It is concluded that the detectors will not work properly when tweets expressing the same meaning are converted into other image styles, which proves that the detectors are not robust enough in this respect.

2) Case Where Images and Texts Do Not Correspond: Implementation Details: In this section, we randomly scrambled the images corresponding to the tweets in the test set to express the inconsistency of the images and texts. The experiment process is similar to Section VI-A1. Firstly, the correspondence between images and texts is disrupted, and then put into the model trained from the clean examples. We experiment on Twitter data and BDANN model.

Results and Analysis: The results show that when the content is unchanged, the detectors cannot identify the tweets' authenticity where images and texts do not correspond. This suggests that we should not only pay attention to the performance improvement, but also to the connection between images and texts, such as semantic consistency. Fig. 12(a) shows the performance when the images' style is transferred and the case where images and texts do not correspond.

Answer to RQ3: (1) The visual modality of the multi-modal detectors is less robust, and the detection accuracy of the news containing the adversarial image with the same perturbation ratio drops more ($\epsilon = 0.1$, the adversarial image drops by more than 30%, the adversarial text drops less than 10%); (2) The detector cannot correctly extract the features of the image after style transfer; (3) When the visual and textual information do not match, the detection performance of the detector decreases significantly.

B. Robustness Evaluation of Model Bias

In addition to adversarial attacks and backdoor attacks, we also conduct a bias evaluation on these detectors to evaluate whether the detectors rely on different features differently when making decisions. Inspired by [62] and the above results (the visual features cause greater damage to the detector), it is worth knowing whether the detectors are biased toward a specific feature, such as the visual feature.

When evaluating the text, we replace the real news with the texts of fake news and ensure that the image does not change. Meanwhile, we replace the fake news with the texts of real news. Then we use the models trained on the clean example to test it. The same process when evaluating the image. It's worth noting that when replacing, we do it in the same event, instead of randomly replacing other irrelevant content.

We test the fake category and the real category separately. The results are shown in Fig. 12(b). Regarding the text, no matter what kind of replacement it is, it will not have much impact on the model. However, the replacement of images significantly impacts the model's performance, especially for the fake category. This means that the combination of fake text and the real image seems confusing to the detectors, reducing the accuracy to 6.44%. This also shows that images seem to account for a large proportion of the detector's judgment of fake tweets. This further explains our conclusions in Section V-D and Section V-E: compared with textual features, visual features are more susceptible to adversarial attacks and backdoor attacks, which greatly reduces the detectors' performance. This is because the detectors rely more on visual features, especially when making judgments on fake examples.

TABLE IV
THE DETECTION ACCURACY OF FIVE MULTI-MODAL DETECTOR ON FUZZY NEWS

Detector	Detection Accuracy			
	Original	Fuzzy Text	Fuzzy Image	Fuzzy Both
Att-RNN	0.682	0.678	0.679	0.673
EANN	0.719	0.708	0.693	0.685
MVAE	0.745	0.741	0.739	0.730
BDANN	0.830	0.822	0.817	0.804
SpotFake	0.777	0.771	0.761	0.759

We also explore the bias of multi-modal fake news detectors in benign scenarios. We added random noise (pixels of the image and random letters of the text) to the image and text of the original news to simulate the scenarios where one of the modal information is blurred in news. Specifically, we randomly selected 100 real news and 100 fake news, and added random noise to their text and images respectively. The variation in detection accuracy of the five multi-modal detectors over these 200 examples is reported in Table IV. Experimental results show that these five deep learning-based multi-modal fake news detectors have less modality bias in benign scenarios than in malicious attack scenarios. A small amount of random perturbation ($<$0.1) hardly affects their detection performance. Even if random perturbations are added to the text and image modalities of the input news at the same time, the impact on the detection performance is small. It can be seen that the multi-modal fake news detector based on deep learning is robust against random noise.

TABLE V
THE TIME COST OF DIFFERENT ATTACK METHODS

Detector	Robustness Testing Methods (s)						
	FGSM	DeepFool	PGD	VIPER	HotFlip	BI	BT
Att-RNN	0.1	0.38	0.29	1.72	1.21	12.1	28.6
EANN	0.16	0.61	0.57	2.41	2.11	14.9	34.9
MVAE	0.21	0.65	0.59	3.15	2.98	16.8	40.2
BDANN	0.5	1.45	1.33	7.92	8.01	17.6	48.6
SpotFake	0.25	0.71	0.64	5.13	6.12	17.2	44.3

C. Time Cost of Malicious Attack

We also discuss the time cost of various attack methods to analyze the possibility of these malicious behaviors being implemented in real-world scenarios. The time required to compute one example for adversarial attack methods and backdoor attack methods to perform a backdoor training are shown in Table V. All detectors are trained on Twitter. FGSM, DeepFool, PGD, VIPER, and HotFlip represent three visual modal adversarial attack methods and two textual modal adversarial attack methods, respectively. BI and BT represent the BadNets poisoning attack methods of visual modal and text modal, respectively. For adversarial attacks, we count the average time taken to generate an adversarial image. For poisoning attacks, we count the poisoning training time required to increase the poisoning success rate to more than 50%.

Most robustness testing methods (FGSM, DeepFool, PGD) consume only a small amount of time compared to the time for multi-modal fake news detection. Among them, the robustness testing methods of text modality (VIPER, HotFlip) consumes more time (one test time exceeds one fake news detection time). On average, it only takes 0.5 s to generate a set of multi-modal news with attack effects, which can bypass the detection of these five multi-modal fake news detectors. And the process of adversarial attack can be automatically realized by the machine. The poisoning training of BadNets only takes a small amount of extra time in the process of generating the patch (<0.001 s per example). And on average, only five epochs poisoning trainings are required to achieve a poisoning attack success rate of more than 50%. Malicious developers can covertly implement poisoning attacks in the process of training multi-modal fake news detectors. They are great threats to deep learning-based multi-modal fake news detectors.

D. Writing Styles, Image Forgery and Attacks

Writing style changes and image forgery are two common fake news generation strategies. They can confuse some fake news detectors [63], [64]. There are several important differences between the adversarial attack / poisoning attack methods for images and text mentioned in this work and the writing styles and image forgery methods.

- *Different attack targets:* Adversarial attacks and poisoning attacks were first proposed in the security field of deep learning. They target various deep learning models (feature extractors), while the writing style and image forgery are designed to deceive news readers.
- *Different fake budgets:* Writing style and image forgery methods usually rely on artificially generated fake news because it needs to consider more semantic features. Adversarial attacks and poisoning attacks can be automated through algorithms. These methods generally do not consider the semantic characteristics of examples, but constrain the scale of attacks through disturbance thresholds to achieve concealment purposes.
- *Different attack generality:* Writing style and image forgery methods usually fool some specific fake news detectors, while adversarial attacks and poisoning attacks have general attack capabilities on deep learning based fake news detectors.
- *Relationship between them:* Adversarial attack and backdoor attack methods can be used as optimizations to help fake news produced using writing styles and image forgery methods fool deep learning based detectors.

Answer to RQ4: The detection performance of multi-modal detectors can be improved using simple defense methods: (1) Image resize can improve the robustness of the detector against visual modality attacks imposed by malicious users (the accuracy can be improved by more than 30%); (2) AC defense methods can improve detection robustness to visual modality attacks injected by malicious developers (the accuracy can reach more than 90% of that in clean condition).

VII. Conclusion

This work conducts a comprehensive evaluation of five multi-modal fake news detectors, including adversarial attacks, backdoor attacks, and bias evaluation. The results show that visual features are the common vulnerability of these detectors. We find the reason during the bias evaluation: the detectors rely more on visual features when making decisions, especially when judging fake news, which suggests researchers pay more attention to visual features when they improve the robustness of these detectors, especially the images corresponding to trending events. We also found that both the detection performance and the robustness are positively correlated with the performance of image feature extractors, which provides us with an idea to optimize the detector. In addition, we find that the best-performing model is not necessarily the most robust. Considering the correlation between images and texts is also significantly important to improve the detectors' robustness. Finally, we defend against adversarial attacks and backdoor attacks on the visual features, respectively, which effectively improve the robustness of these detectors. The experiment related data and code are available at https://github.com/kenan976431/Robustness_Multi-modal_Detector.

Our work is a preliminary exploration of these multi-modal fake news detectors' robustness. Several challenges remain, for example, we choose several classic attack and defense methods such as FGSM and image resizing to evaluate these detectors. In future works, we will try confrontation in more complex scenarios and more modal data (such as video and social context) to evaluate the detectors. In addition, fake news is often extremely provocative, leading to its sentiment is often extreme. Therefore, we will also pay more attention to sentiment analysis in fake news detection tasks in future works, which may bring new possibilities to the robustness of these detectors. We only discuss multi-modal fake news detection in offline scenarios, more widely used, robustness analysis on different news publishing platforms and online scenarios detection will be carried out in future work.

REFERENCES

[1] D. Varshney and D. K. Vishwakarma, "A unified approach for detection of Clickbait videos on YouTube using cognitive evidences," *Appl. Intell.*, vol. 51, no. 7, pp. 4214–4235, 2021.

[2] M. Farajtabar et al., "Fake news mitigation via point process based intervention," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1097–1106.

[3] S. V. D. Linden, J. Roozenbeek, and J. Compton, "Inoculating against fake news about COVID-19," *Front. Psychol.*, vol. 11, 2020, Art. no. 2928.

[4] P. Meel and D. K. Vishwakarma, "Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities," *Expert Syst. with Appl.*, vol. 153, 2020 Art. no. 112986.

[5] Y. Wang et al., "EANN: Event adversarial neural networks for multi-modal fake news detection," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 849–857.

[6] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newslett.*, vol. 19, no. 1, pp. 22–36, 2017.

[7] D. Varshney and D. K. Vishwakarma, "Hoax news-inspector: A real-time prediction of fake news using content resemblance over web search results for authenticating the credibility of news articles," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 9, pp. 8961–8974, 2021.

[8] P. Meel and D. K. Vishwakarma, "Machine learned classifiers for trustworthiness assessment of web information contents," in *Proc. IEEE Int. Conf. Comput., Commun., Intell. Syst.*, 2021, pp. 29–35.

[9] P. Mee1 and D. K. Vishwakarma, "A temporal ensembling based semi-supervised ConvNet for the detection of fake news articles," *Expert Syst. Appl.*, vol. 177, 2021, Art. no. 115002.

[10] J. V. Tembhurne, M. M. Almin, and T. Diwan, "MC-DNN: Fake news detection using multi-channel deep neural networks," *Int. J. Semantic Web Inf. Syst.*, vol. 18, no. 1, pp. 1–20, 2022.

[11] D. K. Vishwakarma and C. Jain, "Recent state-of-the-art of fake news detection: A review," in *Proc. IEEE Int. Conf. Emerg. Technol.*, 2020, pp. 1–6.

[12] S. R. Sahoo and B. B. Gupta, "Multiple features based approach for automatic fake news detection on social networks using deep learning," *Appl. Soft Comput.*, vol. 100, 2021, Art. no. 106983.

[13] K. A. Barakat, A. Dabbous, and A. Tarhini, "An empirical approach to understanding users' fake news identification on social media," *Online Inf. Rev.*, vol. 45, pp. 1080–1096, 2021.

[14] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel visual and statistical image features for microblogs news verification," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 598–608, Mar. 2017.

[15] B. Malhotra and D. K. Vishwakarma, "Classification of propagation path and tweets for rumor detection using graphical convolutional networks and transformer based encodings," in *Proc. IEEE 6th Int. Conf. Multimedia Big Data*, 2020, pp. 183–190.

[16] P. Meel and D. K. Vishwakarma, "Deep neural architecture for veracity analysis of multimodal online information," in *Proc. IEEE 11th Int. Conf. Cloud Comput.*, 2021, pp. 7–12.

[17] D. Varshney and D. K. Vishwakarma, "A review on rumour prediction and veracity assessment in online social network," *Expert Syst. Appl.*, vol. 168, 2021, Art. no. 114208.

[18] C. Boididou et al., "Verifying multimedia use at mediaeval 2015," *MediaEval*, vol. 3, no. 3, 2015, Art. no. 7.

[19] S. Eger et al., "Text processing like humans do: Visually attacking and shielding NLP systems," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 1634–1647.

[20] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Representations*, 2015.

[21] S. Ren, Y. Deng, K. He, and W. Che, "Generating natural language adversarial examples through probability weighted word saliency," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1085–1097.

[22] A. Mishra, B. B. Gupta, and R. C. Joshi, "A comparative study of distributed denial of service attacks, intrusion tolerance and mitigation techniques," in *Proc. IEEE Eur. Intell. Secur. Inform. Conf.*, 2011, pp. 286–289.

[23] B. Gupta, S. Gupta, S. Gangwar, M. Kumar, and P. Meena, "Cross-site scripting (XSS) abuse and defense: Exploitation on several testing bed environments and its defense," *J. Inf. Privacy Secur.*, vol. 11, no. 2, pp. 118–136, 2015.

[24] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "BadNets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47230–47244, 2019.

[25] J. Dai, C. Chen, and Y. Li, "A backdoor attack against LSTM-based text classification systems," *IEEE Access*, vol. 7, pp. 138872–138878, 2019.

[26] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy*, 2017, pp. 39–57.

[27] H. Ali et al., "All your fake detector are belong to us: Evaluating adversarial robustness of fake-news detectors under black-box settings," *IEEE Access*, vol. 9, pp. 81678–81692, 2021.

[28] B. Liu and I. Lane, "Attention-based recurrent neural network models for joint intent detection and slot filling," in *Proc. Interspeech*, 2016, pp. 685–689.

[29] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "MVAE: Multimodal variational autoencoder for fake news detection," in *Proc. World Wide Web Conf.*, 2019, pp. 2915–2921.

[30] T. Zhang et al., "BDANN: Bert-based domain adaptation neural network for multi-modal fake news detection," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2020, pp. 1–8.

[31] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, "SpotFake: A multi-modal framework for fake news detection," in *Proc. IEEE 5th Int. Conf. Multimedia Big Data*, 2019, pp. 39–47.

[32] H. Zheng et al., "NeuronFair: Interpretable white-box fairness testing through biased neuron identification," in *Proc. 44th Int. Conf. Softw. Eng.*, 2022, pp. 1519–1531.

[33] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[34] S. Kumar, R. West, and J. Leskovec, "Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes," in *Proc. 25th Int. Conf. World Wide Web*, 2016, pp. 591–602.

[35] C. Song, C. Yang, H. Chen, C. Tu, Z. Liu, and M. Sun, "CED: Credible early detection of social media rumors," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 8, pp. 3035–3047, Aug. 2021.

[36] G. Shrivastava, P. Kumar, R. P. Ojha, P. K. Srivastava, S. Mohan, and G. Srivastava, "Defensive modeling of fake news through online social networks," *IEEE Trans. Comput. Social Syst.*, vol. 7, no. 5, pp. 1159–1167, Oct. 2020.

[37] 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. Multimedia*, 2017, pp. 795–816.

[38] D. K. Vishwakarma, D. Varshney, and A. Yadav, "Detection and veracity analysis of fake news via scrapping and authenticating the web search," *Cogn. Syst. Res.*, vol. 58, pp. 217–229, 2019.

[39] P. Meel and D. K. Vishwakarma, "HAN, image captioning, and forensics ensemble multimodal fake news detection," *Inf. Sci.*, vol. 567, pp. 23–41, 2021.

[40] C. Raj and P. Meel, "ConvNet frameworks for multi-modal fake news detection," *Appl. Intell.*, vol. 51, no. 11, pp. 8132–8148, 2021.

[41] S. Qian, J. Hu, Q. Fang, and C. Xu, "Knowledge-aware multi-modal adaptive graph convolutional networks for fake news detection," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 17, no. 3, pp. 1–23, 2021.

[42] L. Cui, S. Wang, and D. Lee, "SAME: Sentiment-aware multi-modal embedding for detecting fake news," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, 2019, pp. 41–48.

[43] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.

[44] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proc. IEEE Eur. Symp. Secur. Privacy*, 2016, pp. 372–387.

[45] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2574–2582.

[46] J. Ebrahimi, A. Rao, D. Lowd, and D. Dou, "HotFlip: White-box adversarial examples for text classification," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 31–36.

[47] R. Jia and P. Liang, "Adversarial examples for evaluating reading comprehension systems," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 2021–2031.

[48] J. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, and Y. Qi, “TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP,” in *Proc. Conf. Empirical Methods Natural Lang. Process.: Syst. Demonstrations*, 2020, pp. 119–126.

[49] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, “Is bert really robust? a strong baseline for natural language attack on text classification and entailment,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 34 no. 05, 2020, pp. 8018–8025.

[50] B. Wang, B. Pan, X. Li, and B. Li, “Towards evaluating the robustness of chinese bert classifiers,” 2020, *arXiv:2004.03742*.

[51] Y. Liu et al., “Trojaning attack on neural networks,” in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2018.

[52] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, “How to backdoor federated learning,” in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 2938–2948.

[53] K. Kurita, P. Michel, and G. Neubig, “Weight poisoning attacks on pretrained models,” in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2793–2806.

[54] Z. Sun, P. Kairouz, A. T. Suresh, and H. B. McMahan, “Can you really backdoor federated learning ?,” 2019, *arXiv:1911.07963*.

[55] M. Du, F. Yang, N. Zou, and X. Hu, “Fairness in deep learning: A computational perspective,” *IEEE Intell. Syst.*, vol. 36, no. 4, pp. 25–34, Jul./Aug. 2021.

[56] G. Joshi, R. Walambe, and K. Kotecha, “A review on explainability in multimodal deep neural nets,” *IEEE Access*, vol. 9, pp. 59800–59821, 2021.

[57] I. Gat, I. Schwartz, A. Schwing, and T. Hazan, “Removing bias in multimodal classifiers: Regularization by maximizing functional entropies,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 3197–3208.

[58] Y. Guo et al., “On modality bias recognition and reduction,” *ACM Trans. Multimedia Comput., Commun., Appl.*, 2022.

[59] H. Zheng, J. Chen, H. Du, W. Zhu, S. Ji, and X. Zhang, “GRIP-GAN: An attack-free defense through general robust inverse perturbation,” *IEEE Trans. Dependable Secure Comput.*, vol. 19, no. 6, pp. 4204–4224, Nov./Dec. 2022.

[60] G. Hinton and L. van der Maaten, “Visualizing data using T-Sne,” *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605,, Nov. 2008.

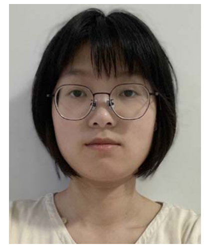
[61] B. Chen et al., “Detecting backdoor attacks on deep neural networks by activation clustering,” *AAAI Workshop*, 2019.

[62] Y. Li et al., “Shape-texture debiased neural network training,” in *Proc. Int. Conf. Learn. Representations*, 2020.

[63] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, “A stylometric inquiry into hyperpartisan and fake news,” in *Proc. 56th Ann. Meeting Assoc. Computat. Linguist.*, vol. 1, pp. 231–240, 2018.

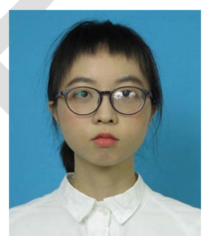
[64] M. Huh, A. Liu, A. Owens, and A. A. Efros, “Fighting fake news: Image splice detection via learned self-consistency,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 101–117.

Jinyin Chen (Member, IEEE) received the B.S. and Ph.D. degrees from the Zhejiang University of Technology, Hangzhou, China, in 2004 and 2009, respectively. She studied evolutionary computing with the Ashikaga Institute of Technology, Ashikaga, Japan, in 2005 and 2006. She is currently a Professor with the Zhejiang University of Technology. Her research interests include artificial intelligence security, graph data mining and evolutionary computing.

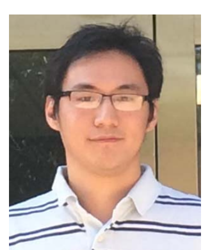


Chengyu Jia received the bachelor’s degree in 2020 from the Zhejiang University of Technology, Hangzhou, China, where she is currently working toward the graduation degree with the College of Information Engineering. Her research interests include deep learning, artificial intelligence, and adversarial attack and defense.

Haibin Zheng received the bachelor’s degree from the Zhejiang University of Technology, Hangzhou, China, in 2017. He is currently working toward the Ph.D. degree with the College of Information Engineering, Zhejiang University of Technology. His research interests include deep learning, artificial intelligence, and adversarial attack and defense.



Ruoxi Chen received the bachelor’s degree in 2020 from the Zhejiang University of Technology, Hangzhou, China, where she is currently working toward the graduation degree with the College of Information Engineering. Her research interests include deep learning, artificial intelligence, and adversarial attack and defense.



Chenbo Fu (Member, IEEE) received the B.S. degree in physics from the Zhejiang University of Technology, Hangzhou, China, in 2007, and the M.S. and Ph.D. degrees in physics from Zhejiang University, Hangzhou, in 2009 and 2013, respectively. He was a Postdoctoral Researcher with the College of Information Engineering, Zhejiang University of Technology. He was a Visiting Scholar with the Department of Computer Science, University of California at Davis, Davis, CA, USA, in 2014. In 2017, he was a Research Assistant with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong. He is currently a Lecturer with the Institute of Cyberspace Security, Zhejiang University of Technology. His research interests include network-based algorithms design, social networks, data mining, chaos synchronization, network dynamics, and machine learning.