Applied Artificial Intelligence

An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

Robust Rumor Detection based on Multi-Defense Model Ensemble

Fan Yang & Shaomei Li

To cite this article: Fan Yang & Shaomei Li (2023) Robust Rumor Detection based on Multi-Defense Model Ensemble, Applied Artificial Intelligence, 37:1, 2151174, DOI: 10.1080/08839514.2022.2151174

To link to this article: https://doi.org/10.1080/08839514.2022.2151174

Published online: 28 Dec 2022.

APPLIED ARTIFICIAL INTELLIGENCE
2023, VOL. 17, NO. 1, e2151174 (125 pages)
https://doi.org/10.1080/08839514.2022.2151174

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Robust Rumor Detection based on Multi-Defense Model Ensemble

Fan Yang and Shaomei Li

Institute of information technology, Information Engineering University, Zhengzhou, People's Republic of China

ABSTRACT

The development of adversarial technology, represented by adversarial text, has brought new challenges to rumor detection based on deep learning. In order to improve the robustness of rumor detection models under adversarial conditions, we propose a robust detection method based on the ensemble of multi-defense model on the basis of several mainstream defense methods such as data enhancement, random smoothing, and adversarial training. First, multiple robust detection models are trained based on different defense principles; then, two different ensemble strategies are used to integrate the above models, and the detection effect under different ensemble strategies is studied. The test results on the open-source dataset Twitter15 show that the proposed method is able to compensate for the shortcomings of a single model by ensembling different decision boundaries to effectively defend against mainstream adversarial text attacks and improve the robustness of rumor detection models compared to existing defense methods.

ARTICLE HISTORY
Received 22 June 2022
Revised 11 November 2022
Accepted 18 November 2022

Introduction

Rumor detection technology is a sub-task of text classification in the natural language processing field, judging the authenticity of a message by identifying input text and other characteristics (Shu et al. 2017). With the development of deep learning technology and its popularization and application in the field of natural language processing, the rumor detection model based on deep learning has greatly improved the accuracy of rumor detection and become a mainstream method. This type of method mainly regards rumor detection as a text classification task and applies deep neural network models to make a high-level representation of the input text and classify it (Gao, Liang, and Jiang et al. 2020). However, with the wide application of rumor detection technology in the real world, some criminals use the fragility of deep neural networks to try to deceive rumor detection models, achieve the purpose of

CONTACT Fan Yang, le2jh@foxmail.com Institute of information technology, Information Engineering University, Zhengzhou, People's Republic of China

circumventing supervision, and bring new adjustments to rumor detection technology.

Adversarial text is the current mainstream adversarial method that deceives target models by adding perturbations to characters, words, or sentences. Related studies (Cheng et al. 2020a; Goodfellow, Shlens, and Szegedy 2015; Ling, Ji, and Zou et al. 2019) show that natural language processing models based on deep neural networks exhibit great vulnerability to maliciously generated adversarial text. As (Zhou, Guan, and Bhat et al. 2019) noted, tampering with words or characters in news text content may mislead the detector into detecting rumors as real news. To avoid attacks being perceived by humans, attackers typically use synonyms to generate adversarial text (Jin et al. 2020; Li, Ji, and Du et al. 2019; Ren, Deng, and He et al. 2019), circumventing the defense methods used in the detection, such as automatic spelling and grammar checking, while preserving the original semantic information well.

Given the above synonym substitution-type attack problem, existing researches have proposed defense methods based on data enhancement (Si, Zhang, and Qi et al. 2020; Wang and Bansal 2018) and adversarial training (Madry, Makelov, and Schmidt et al. 2018; Miyato, Dai, and Goodfellow 2017; Zhu, Cheng, and Gan et al. 2019) to enhance the robustness of English text classification models. These methods can also be migrated to the rumor detection model. The key idea of the former approach is to add manual rule-making adversarial text to the training set to assist classifier training, but it is only for specific types of attacks and is difficult to cover multiple types of attacks. However, real-world attacks on text input tend to be ever-changing and the search space against text is growing exponentially. The adversarial training-based approach introduces the defense idea of minimum-maximum optimization in the image field, improves the regularization ability of the model by adding norm-bounded interference to the word embedding, expands the decision boundary and enhances the robustness of the model. There have also been studies (Gupta et al. 2022) using machine translation, which translates input text from the source language to the target language and translates it back into the source language again before feeding it to the classifier, but this method has a large semantic loss.

To improve the robustness of rumor detection models under adversarial conditions, we study the defense effectiveness of current mainstream adversarial text defense methods and propose a defense method based on model ensemble, which further enhances the success rate of rumor detection models in dealing with adversarial texts by setting a reasonable ensemble strategy to compensate for the decision failure of a single robust model in the face of adversarial texts. Specifically, there are four main points of innovation in our approach:

(1) Based on the idea of data enhancement, with reference to the current mainstream adversarial text generation method, synonyms are selected from the two large synonyms knowledge bases of Hownet and WordNet to replace the important words in the original text. The training set is expanded by artificially generating adversarial text through this method, to train a robust rumor detection model 1;
(2) Extending the idea of random smoothing in the field of image processing to the discrete and structured space of the text and training a robust rumor detection model 2 through random scrambling;
(3) Using the standard adversarial training method PGD to train a robust rumor detection model 3;
(4) In the detection stage, the model ensemble idea is adopted to integrate the results of rumor detection models 1, 2 and 3 to further improve the confrontation and defense effect of the model.

Related Work

Rumor Detection Classifier

Rumor detection is essentially a text classification problem, with inputs being a sequence of text words and output being a single label. Current rumor detection models based on deep learning have achieved good results in the open-source rumor detection dataset, indicating that rumor detection based on deep neural network extraction of text features is effective (Gao, Liang, and Jiang et al. 2020).

In general, convolutional neural network (CNN) can be used to extract text semantic features (Yuan, Ma, and Zhou et al. 2019), recurrent neural network (RNN) and its variants GRU, LSTM, etc. are used to model the dependencies between text and forwarding sequences (Ma et al. 2016; Ruchansky, Seo, and Liu 2017; Shu, Cui, and Wang et al. 2019), and studies (Shu, Cui, and Wang et al. 2019; Yuan, Ma, and Zhou et al. 2019) have also introduced attention networks to further extract the deep features of message text. Most of the existing rumor detection classifiers integrate neural networks with different structures to model and classify rumors themselves and their propagation processes end-to-end.

Adversarial Text Generation

Traditional methods of adversarial text generation in the NLP domain include character-level (Eger, Şahin, and Rücklé et al. 2019; He, Lyu, and Xu et al. 2021) substitution of similar letters, addition of symbols between characters, and word-level (Jin et al. 2020; Li, Ji, and Du et al. 2019; Ren, Deng, and He et al. 2019) synonym substitution. However, the former can be easily corrected by spell checking (Hládek, Staš, and Pleva 2020), while most of the adversarial text faced in rumor detection tasks is handmade by netizens, who prefer to adopt a word variation strategy, that is, modify certain important words without affecting semantics. (Ren, Deng, and He et al. 2019) use WordNet as[1] a thesaurus to replace the generated adversarial text, and (Zang, Qi, and Yang et al. 2019) further introduces the Hownet [2]knowledge base to expand the search scope of synonyms, not only increases the sample size but also makes the generated adversarial text closer to the real semantic information.

Defense Method Against Synonym Substitution Type Attacks

(Wang and Yang 2015) first attempted to add interfered text to the training to improve the robustness of the model, but due to the low efficiency of adversarial text generation, the robustness performance of the model was limited. Methods based on random smoothing(Ye, Gong, and Liu 2020) input text training by constructing random sets and using the statistical properties of sets to prove robustness. (Madry, Makelov, and Schmidt et al. 2018; Zhu, Cheng, and Gan et al. 2019) propose adversarial training methods such as PGD and FreeLB based on the minimum-maximum optimization formula. The study (Wang, Tang, and Lou et al. 2021) proposes a privacy framework wordDP based on an exponential mechanism, which applies differential privacy methods to robustly verify synonym substitution attacks in text classification to ensure that small changes in input do not lead to sharp changes in output. (Li, Song, and Zeng et al. 2022) proposes a rebuild-ensemble framework that reconstructs text using the mask-fill capability of pre-trained models and uses these texts with less adversarial effects for predictions for better robustness.

Method Based on Multi-Defense Model Ensemble

As shown in Figure 1, the adversarial text successfully deceives the original detection model f^{Original} without adopting a defense strategy, misleading the detection model to give “non-rumor” error results. In order to improve the robustness, we first use three defense methods to improve the detection model f^{Original}, and obtain a rumor detection model based on data enhancement f^{Data}, a rumor detection model based on random smoothing f^{RS}, and a rumor detection model based on adversarial training f^{PGD}.

In the detection process, the above three robust rumor detection models are used to detect the input text, and then the detection results of the three models are integrated to further improve the effectiveness of adversarial defense.

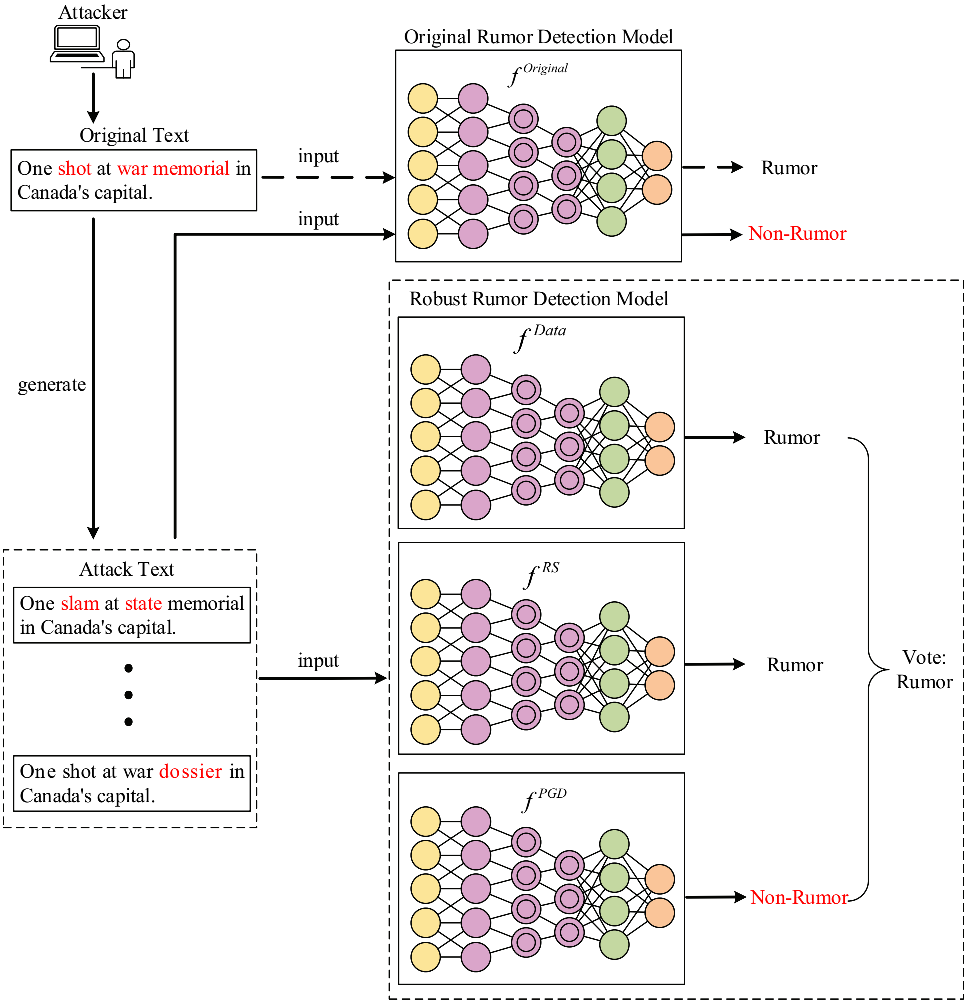
Attacker
Original Text
One shot at war memorial in Canada's capital.
input
input
generate
Original Rumor Detection Model
$f^{Original}$
Rumor
Non-Rumor
Robust Rumor Detection Model
f^{Data}
Rumor
f^{RS}
Rumor
f^{PGD}
Non-Rumor
Vote: Rumor
Attack Text
One slam at state memorial in Canada's capital.
One shot at war dossier in Canada's capital.
input

Figure 1. Rumor detection framework based on multi-defense model ensemble.

The detection model and ensemble strategy under the three single defense strategies are introduced below.

Data Augmentation Based Robust Detection Model

In real-world adversarial text scenarios, attackers often locate important words by changes in confidence information because they cannot obtain specific gradient information about the classifier. This paper extends the training set by mimicking the attacker’s method of generating adversarial text, and improves the robustness of the model through data augmentation. Given an input sentence sequence $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$, where x_i represents the $\mathrm{i} - \mathrm{th}$ word, we use the scoring function to determine the importance of the $\mathrm{j} - \mathrm{th}$ word in $\mathbf{x}$:

$$C_{x_j} = f_y(x_1, x_2, \ldots, x_n) - f_y(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n) \tag{1}$$

In formula (1), f_y is the classifier model, and C_{x_j}calculates the difference in confidence between the two-text classification results before and after the word x_j is deleted from the original text. Using the above method, the confidence difference of each word is calculated in turn and sorted according to the size of the difference. The larger the confidence difference, the higher the importance score.

At the same time, the words are replaced by the synonyms in WordNet and Hownet until the predicted value of the classifier changes, so far two adversarial texts are generated. WordNet is a large, hand-organized semantic dictionary in which synonyms are grouped into synonyms. Hownet is a knowledge base of sememes, and in general, words with the same sememes are represented by the same meaning and can be substituted for each other. To ensure that the semantics do not change after substitution, we have made a provision that the average revision rate of sentences does not exceed 20%.

Through the above method, the adversarial text data of twice the size of the original training set can be generated, and the adversarial text generated by the external knowledge base can be directly added to the original training set. If the rumor detection model can be trained at the same time, then a robust model f^{Data} based on data enhancement can be obtained.

Random Smoothing Based Robust Detection Model

The definition of a smoothing model in SAFER (Ye, Gong, and Liu 2020) is as follows:

$$f^{RS}(\mathbf{x}) = \arg\max_{c \in y} P_{\mathbf{z} \sim \prod_{\mathbf{x}}} (f(\mathbf{z}) = c) \tag{2}$$

As shown in Equation (2), what the robust detection model f^{RS} after random smoothing needs to be satisfied is that when the sentence$\mathbf{x}$in the original input text adds random perturbation to $\mathbf{z}$, the model predicts that $\mathbf{z}$still belongs to the original category c.

Unlike the data augmentation-based defense method in Section 3.1, which directly transforms the input text, the random smoothing-based defense method used in this section transforms the word embedding representation of the input text at the embedding layer of the model to allow the model to learn more adversarial forms. Therefore, unlike the substitution of the original text in section 3.1 based on the dictionary or knowledge base, this section constructs a perturbation set P_x in the context-aware word vector space, that is, the embedding of the original word in the text is replaced by the K nearest neighbor embedding. Drawing on existing research (Ye, Gong, and Liu 2020), we used the Glove model for word embedding and set K to 10.

For a sentence $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ in the input text, its perturbation distribution $\prod_{\mathbf{x}}$ is defined by randomly perturbing each word x_i as a word in the perturbation set P_x with the same probability, the formula expressed as:

$$\prod_{\mathbf{x}}(\mathbf{z}) = \prod_{i=1}^{n} \frac{II\{\mathrm{z_i} \in P_{x_i}\}}{|P_{x_i}|} \tag{3}$$

where $\mathbf{z} = \{z_1, z_2, \ldots, z_n\}$ is the sentence after the perturbation, $|P_{x_i}|$ is the size of the word $x_i$perturbation set, and II is the indicative function.

The smooth representation of the original text embedding obtained based on the above method is sent into the rumor detection model to train, which can expand the data distribution exposed by the model, so as to obtain a more robust detection model f^{RS}.

Adversarial Training Based Robust Detection Model

Sections 3.1 and 3.2 augment the data that can be used to train the model by means of raw text transformation and text embedding layer transformation, respectively, to improve the robustness of the model, and this section adopts PGD (Madry, Makelov, and Schmidt et al. 2018), a defense method based on adversarial training.

The principle of PGD is to minimize model parameters within a range to resist worst-case perturbations, as shown in the formula:

$$\min_{\theta} E_{(x,y)\sim D}[\max_{\|\delta\| \le \varepsilon} L(f_\theta(\mathbf{X} + \delta), y)] \tag{4}$$

where D is the distribution of inputs,$\mathbf{X}$is the embedded representation of the input sentence$\mathbf{X}$,yis the classification label, and L is the loss function of the classifier, whose parameter distribution is expressed as θ. To solve the internal maximization problem, PGD employs a gradient projection descent algorithm:

$$\delta_{t+1} = \prod_{\|\delta\|_F \le \varepsilon} (\delta_t + \alpha \frac{g(\delta_t)}{\|g(\delta_t)\|_F}) \tag{5}$$

where $g(\delta_t) = \quad {}_{\delta}L(f_\theta(\mathbf{X} + \delta), y)$ is the gradient of the loss function L relative to the perturbation δ, $\prod_{\delta_F \le \varepsilon}$ represents the projection on the εnorm, and t finds the ascending step of the "worst-case" perturbation δ with a step α.

In the process of rumor detection model training, after iterating on K times δ by PGD algorithm, the model parameters are updated by taking the gradient of the last perturbation, so as to obtain a robust detection model f^{PGD} after adversarial training.

Multi-Defense Model Ensemble Strategy

For the three robust models generated by the above different methods, this section mainly examines the ensemble effects under two ensemble strategies: logits-summed and majority-vote (Cheng et al. 2020b). Logits refer to the input of the softmax function. The former is a direct average of the logits generated by each classifier, and the latter is to count the detection results of each classifier and use the voting results as the output of the rumor detection.

We define the input as $(\mathbf{x}, y)$, where $\mathbf{x}$ is the sequence of input words and $y \in [C]$ is the classification label. The logits of the $\mathrm{i}-\mathrm{th}$ classifier are defined by $f_{\mathrm{i}}(\mathbf{x})$,$i \in [1, 2, 3]$, and the label predicted by each classifier is:

$$F_{\mathrm{i}}(\mathbf{x}) = \underset{c=0,\ldots,C-1}{\arg\max} f_i(\mathbf{x})_c \tag{6}$$

The output of the integrated classifier with the logits-summed strategy is $f(\mathbf{x}) = \frac{\sum_{i=1}^{3} f_i(\mathbf{x})}{3}$, and the prediction label is:

$$F(\mathbf{x}) = \underset{c=0,\ldots,C-1}{\arg\max} f(\mathbf{x})_c \tag{7}$$

The integrated classifier prediction labels that use the majority-vote strategy are:

$$F(\mathbf{x}) = \underset{c=0,\ldots,C-1}{\arg\max} \sum_{i=1}^{3} F_i(\mathbf{x}) \tag{8}$$

Through the above strategies, the detection effect of multiple models can be integrated to enhance the defense ability of rumor detection models in the face of unknown attacks.

Experiments

Datasets and Rumor Detection Models

The dataset adopts Twitter15, a classic dataset collected from Twitter, the most popular social media site in the United States, with tweets averaging about 15 words in length and containing four labels, “False Rumor” (FR), “True Tumor” (TR), “Unverified” (UR), and “Non-Rumor” (NR).

The rumor detection model adopts CSI (Ruchansky, Seo, and Liu 2017), Defend (Shu, Cui, and Wang et al. 2019) and GLAN (Yuan, Ma, and Zhou et al. 2019) and Bert (Devlin, Chang, and Lee et al. 2019), the above four models, respectively, use the current mainstream CNN, RNN and attention mechanism in the field of natural language processing to extract deep semantic information for rumor detection, and the detection method and detection effect are representative.

CSI uses LSTM to extract the eigenvectors of the time-series text input, combine the user history information, and apply the full connection layer to classify the output eigenvectors. dEFEND first uses GRU to encode the words and stitches them together to get a word representation that combines the context, because each word contributes differently to the sentence, then the weight of each word is learned through the attention mechanism, and the sentence representation is weighted. The sentence representation is then entered again into the bidirectional GRU extraction context for the timing feature representation for rumor detection. Global-local attention network (GLAN) achieves accurate detection of rumors by combining the text content of rumors with the local semantic information and global structural information in the process of dissemination. First, the input sentence is converted into a vector form, and the CNN is applied on the word vector matrix to extract features; then, using the same method it can get a feature representation of the forwarded text; finally, a multi-head attention mechanism is applied to integrate the features of the original text and forward them into a more advanced semantic representation. The Bert-based rumor detection model represents a detection method using a large-scale pre-trained model.

Attack Methods

In order to evaluate the performance of the model in the face of different attacks in real-world scenarios, we use two black box attack methods: TextFooler and PWWS. In order to avoid being easily detected by the human eye, the average modified word is limited to no more than 20%.

- TextFooler (Jin et al. 2020): First measure the impact of words on the classification results and sort them, then construct candidate substitution words based on counter-fitting word vectors, and select the words that change the target label to replace them.
- PWWS (Ren, Deng, and He et al. 2019): Sort all words based on probability-weighted word significance scores, and then greedily traverse the candidate substitution words until the labels of the model change. We use two synonymous thesauruses, Hownet and WordNet, respectively, to generate candidate replacement words.

Table 1 shows examples of adversarial text generated by the two attack methods. It can be seen that in the case of the same attack algorithm, the main reason for the difference between the adversarial texts lies in different sets of external synonyms.

Table 1. Adversarial Text Examples.

Attack Methods	Original Text	Adversarial Text
TextFooler	North korea may be preparing for a bullet launch.	North korea may be develop for a rocket launch.
PWWS(Wordnet)	North korea may be preparing for a bullet launch.	North korea may be preparing for a projectile launch.
PWWS(Hownet)	North korea may be preparing for a bullet launch.	North korea may be preparing for a cartridge launch.

Evaluation Indicators

We use the prevailing metric accuracy rate (Acc) in rumor detection, while introducing both the attack success rate (Suc) and the average number of times the attacker query model (Que).

- Acc: The number of texts correctly detected by the model divided by the number of all texts;
- Suc: The number of adversarial texts that successfully interfered with the model divided by the number of all adversarial texts;
- Que: A classic metric for evaluating robustness, the higher the average number of times an attacker queries a model, the harder the model is to attack.

According to the definition of the above indicators, the criteria of the robust rumor detection model are high accuracy, low attack success rate, and high queries.

Results and Analysis

In order to construct robust defensive detection methods, this section tests the effects of the four common rumor detection models (f^{Original}) described in Section 4.1 and their improved models under section 3.1 (f^{Data}), 3.2 (f^{RS}) and 3.3 (f^{PGD}) defense methods. The test data adopt the Twitter15 test set in Section 4.1 and use the methods in Section 4.2 to attack it, and the results are shown in Table 2.

Rumor Detection Model Vulnerability Analysis

From the experimental results of the attack based on the original model in the first row of Table 2, the adversarial rumor text causes a significant decrease in the detection accuracy of all four types of rumor detection models, with the CSI model showing the largest decrease and the highest success rate of the attack. This is because the CSI model has the simplest structure and uses only RNN to extract the propagation timing features, which has the weakest robustness. The dEFEND model, on the other hand, has stronger robustness

Table 2. Results of different models on the Twitter15 dataset and its adversarial text.

		Original data	TextFooler			PWWS(Wordnet)			PWWS(Hownet)		
Models		Acc	Acc	Suc	Que	Acc	Suc	Que	Acc	Suc	Que
$f^{Original}$	CSI	0.712	0.207	0.79	36.8	0.166	**0.80**	102.5	0.192	0.82	269.1
	dEFEND	0.771	0.281	0.72	56.9	0.202	0.86	131.8	0.198	0.79	341.7
	GLAN	**0.827**	0.374	0.59	53.4	0.259	0.82	213.8	0.287	**0.71**	322.6
	Bert	0.824	**0.396**	**0.48**	**65.5**	**0.274**	0.81	**296.4**	**0.292**	0.77	**406.0**
f^{Data}	CSI	0.709	0.337	0.74	66.4	0.522	0.64	284.9	0.543	0.61	404.5
	dEFEND	0.717	0.402	0.63	80.4	0.693	0.60	298.5	0.690	0.56	474.8
	GLAN	0.821	0.519	0.53	86.3	0.734	0.57	384.7	**0.795**	**0.47**	487.4
	Bert	**0.824**	**0.522**	**0.45**	**103.9**	**0.791**	**0.52**	**446.7**	0.788	0.49	**579.2**
f^{RS}	CSI	0.704	0.421	0.68	72.1	0.427	0.68	197.3	0.467	0.67	329.4
	dEFEND	0.765	0.497	0.57	87.7	0.519	0.62	294.5	0.518	0.65	382.2
	GLAN	**0.824**	0.613	0.44	98.0	0.633	0.60	337.4	0.620	**0.59**	395.8
	Bert	0.822	**0.624**	**0.37**	**99.2**	**0.674**	**0.54**	**385.3**	**0.623**	0.60	**476.9**
f^{PGD}	CSI	0.719	0.334	0.74	57.9	0.250	0.71	178.5	0.239	0.74	309.3
	dEFEND	0.773	0.412	0.62	61.2	0.362	0.70	204.7	0.351	0.74	365.8
	GLAN	**0.831**	0.470	0.48	74.2	0.417	**0.69**	278.4	0.447	0.63	335.9
	Bert	0.827	**0.488**	**0.39**	**81.8**	**0.494**	0.72	**332.3**	**0.462**	**0.68**	**465.3**

compared to the CSI model because it uses a multi-layer attention mechanism to deeply integrate text semantic and comment information. The GLAN model obtains a more robust node representation by applying the graph attention mechanism, while effectively fusing multidimensional features for detection, reducing the sensitivity of the model to adversarial text and providing better robustness. BERT is the most robust rumor detection model due to its natural use of bidirectional Transformer structure, which can capture deeper semantic information for classification.

Effectiveness of Defense Methods

To facilitate the analysis of the results, the results of Table 2 are classified as a histogram, as shown in Figure 2. The abscissa represents three types of defense methods, the four colors represent the four types of rumor detection models introduced in Section 4.1, and the ordinate coordinate represents the improvement value of the robustness evaluation index, which is analyzed as follows: The results of the three attack methods described in Section 4.2 are shown in Figure 2(a-c), respectively.

Figure 2 shows the improvement effect of the robust performance of the four models under the three types of attack methods, that is, in the case of adversarial text attack, compared with the original model, the increase in the recognition accuracy rate, the decrease in the attack success rate, and the increase in the number of queries of the robust model using the defense method.

It can be seen that all rumor detection models have improved their robustness after applying the three types of defense methods proposed in sections 3.1-3.3, which shows that the three defense methods we have adopted are effective on all rumor detection models. As shown in the first column in the figure, among the four types of models, the CSI model ranked the worst in terms of detection

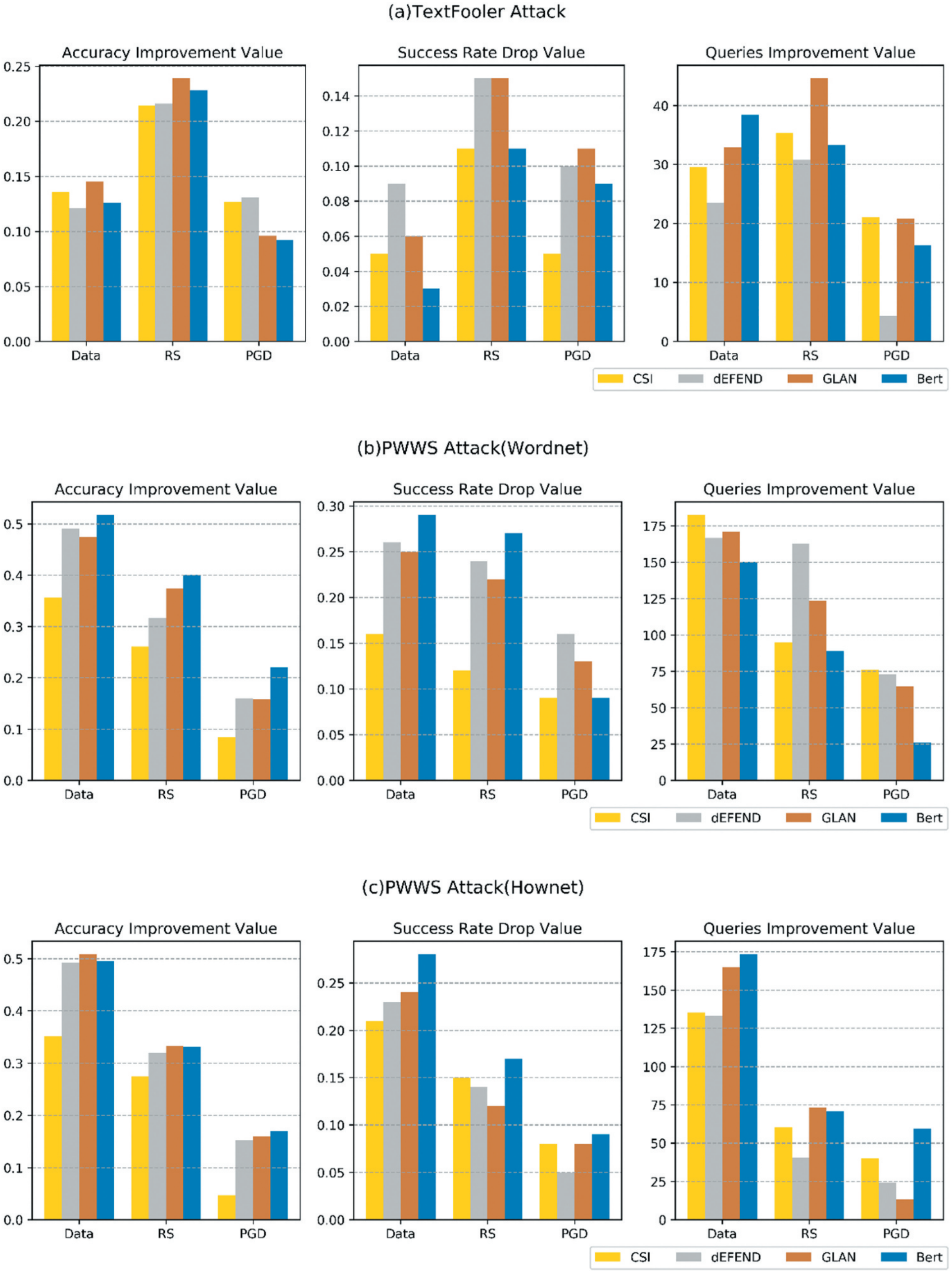

Figure 2. Robust indicators improvement value of each rumor detection model under adversarial conditions.

accuracy improvement after applying the defense method, among which the method based on adversarial training improved the accuracy rate by less than 10%, and the robustness improvement effect was limited. Our guess is that, because the CSI-X neural network structure is relatively simple, it is still vulnerable to attack even if it is defended. While dEFEND and GLAN apply attention mechanisms, Bert uses a bidirectional transformer structure that can extract deep semantic information for rumor detection, thereby reducing the sensitivity of the model. These three types of models have significantly improved their robust performance after applying defense methods.

Complementarity of Defense Methods

From Figure 2, the performance of robust models based on three types of defense methods in the face of different types of attacks can be further analyzed. When faced with TextFooler attacks using counter-fitting word vector exchanges, the robust model based on random smoothing structure has the best defense performance, and the detection accuracy on all four models is improved by more than 20%.

At the same time, the experimental results show that the data enhancement method based on WordNet and Hownet knowledge base has achieved the best defense effect in dealing with both attacks of PWWS, because the training process of the robust model shares the same thesaurus with the attack process of PWWS, which is equivalent to an "open-book exam."

In the face of TextFooler attacks with different substitution strategies, the data enhancement method is average, but it is still superior to the traditional adversarial training method PGD. In other words, PGD was the worst defense against three types of attacks, which also coincided with previous experiments on adversarial training in other text classifications (Madry, Makelov, and Schmidt et al. 2018). The principle of traditional adversarial training is to add small perturbations to the embedding layer to expand the decision boundary, but the embedding vector after the perturbation may not necessarily match the original embedding vector table, so that the perturbation of the embedding layer cannot correspond to the real text input, which is inconsistent with the actual attack scenario. Therefore, the PGD method has limited ability to improve the robustness of the model, and is more as a regularization method to improve the regularization ability of the rumor detection model.

In summary, when faced with a thesaurus attack based on the unknown, we tend to choose a random and smooth approach to defense.

Ensemble Strategy Effectiveness Analysis

Through the analysis of sections 5.2 and 5.3, it can be seen that the robust performance of the model in the face of adversarial text can be improved to varying degrees using a variety of defense methods. From the perspective of multi-model ensembles, this section studies the application effects of the two ensemble strategies in Section 3.4, and the experimental results are shown in Tables 3–6.

Table 3. Experimental results after CSI apply the ensemble strategy.

	Original data	TextFooler		PWWS(Wordnet)		PWWS(Hownet)	
	Acc	Acc	Suc	Acc	Suc	Acc	Suc
$f^{Original}$	0.712	0.207	0.79	0.166	0.80	0.192	0.82
f^{Data}	0.709	0.337	0.74	0.522	0.64	0.543	0.61
f^{RS}	0.704	0.421	0.68	0.427	0.68	0.467	0.67
f^{PGD}	0.719	0.334	0.74	0.25	0.71	0.239	0.74
loggits-summed	0.718	0.425	0.68	0.479	0.62	0.457	0.67
majority-vote	**0.730**	**0.478**	**0.62**	**0.536**	**0.59**	**0.548**	**0.56**

Table 4. Experimental results after dEFEND applies the ensemble strategy.

	Original data	TextFooler		PWWS(Wordnet)		PWWS(Hownet)	
	Acc	Acc	Suc	Acc	Suc	Acc	Suc
$f^{Original}$	0.771	0.281	0.72	0.202	0.86	0.198	0.79
f^{Data}	0.717	0.402	0.63	0.693	0.60	0.69	0.56
f^{RS}	0.765	0.497	0.57	0.519	0.62	0.518	0.65
f^{PGD}	0.773	0.412	0.62	0.362	0.7	0.351	0.74
loggits-summed	0.765	0.464	0.58	0.697	0.60	0.622	0.59
majority-vote	**0.792**	**0.510**	**0.53**	**0.712**	**0.57**	**0.702**	**0.58**

Table 5. Experimental results after GLAN are applied to the ensemble strategy.

	Original data	TextFooler		PWWS(Wordnet)		PWWS(Hownet)	
	Acc	Acc	Suc	Acc	Suc	Acc	Suc
$f^{Original}$	0.827	0.374	0.59	0.259	0.82	0.287	0.71
f^{Data}	0.821	0.519	0.53	0.734	0.57	0.795	0.47
f^{RS}	0.824	0.613	0.44	0.633	0.6	0.62	0.59
f^{PGD}	0.831	0.47	0.48	0.417	0.69	0.447	0.63
loggits-summed	0.789	0.616	0.44	0.747	0.33	0.764	**0.30**
majority-vote	**0.845**	**0.665**	**0.42**	**0.777**	**0.30**	**0.783**	**0.30**

Table 6. Experimental results after Bert applies the ensemble strategy.

	Original data	TextFooler		PWWS(Wordnet)		PWWS(Hownet)	
	Acc	Acc	Suc	Acc	Suc	Acc	Suc
$f^{Original}$	0.824	0.396	0.48	0.274	0.81	0.292	0.77
f^{Data}	0.824	0.522	0.45	0.791	0.52	0.788	0.49
f^{RS}	0.822	0.624	0.37	0.674	0.54	0.623	0.6
f^{PGD}	0.827	0.488	0.39	0.494	0.72	0.462	0.68
loggits-summed	0.827	0.619	0.37	0.780	0.30	0.790	0.29
majority-vote	**0.863**	**0.683**	**0.35**	**0.793**	**0.29**	**0.824**	**0.27**

Judging from the detection results on the original dataset in Tables 3–6 after applying the data enhancement and random smoothing defense methods, the detection accuracy of all rumor detection models on unscrambled clean text was slightly reduced, but the detection accuracy of clean text by the robust model based on adversarial training was improved. That is to say, while the three defense methods we use improve the detection performance of adversarial text, the impact on the detection performance of unscrambled clean text is relatively weak.

At the same time, the loggits-summed ensemble strategy only averages the output results of the model, which weakens the overall diversity of the integration to a certain extent, so the performance improvement of robustness is limited, and sometimes the accuracy of detection is not as high as that of a single model. The voting-based strategy performs best on both clean text and adversarial text. Note that the voting-based ensemble strategy can effectively resist three types of attacks based on different thesauruses and improve the robustness of the model. This is because individual models trained on different loss functions have different decision boundaries, and when a set is formed, it leads to more diversity, thus compensating for the deficiencies between each other. Therefore, the detection results of multiple robust models can be made through voting strategies for the integration of decision-making, which can improve the robustness of the model.

Conclusion

In order to enhance the robustness of the rumor detection model against maliciously produced adversarial text in reality, this paper proposes a rumor detection adversarial defense method based on the ensemble of multiple defense models. This method applies mainstream defense strategies such as data augmentation, random smoothing, and adversarial training to compensate for the shortcomings of a single model by ensembling different model decision boundaries, thus effectively defending against mainstream adversarial text attacks and achieving more robust rumor detection. Through experimental evaluation of the open-source rumor dataset, we prove that the proposed method can effectively improve the effectiveness of rumor detection under adversarial conditions.

Notes

1. https://wordnet.princeton.edu.
2. https://openhownet.thunlp.org.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Cheng, M., J. Yi, P. Y. Chen, H. Zhang, and C.-J. Hsieh. 2020a. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (4):3601–08. doi:10.1609/aaai.v34i04.5767.

Cheng, M., C. J. Hsieh, and I. Dhillon. 2020b. Voting based ensemble improves robustness of defensive models[j]. *arXiv preprint arXiv:2011 14031*.

Devlin, J., M. W. Chang, K. Lee, and Toutanova K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding[c]. *NAACL-HLT* 4171–86.

Eger, S., G. G. Şahin, A. Rücklé, JU Lee, C Schulz, M Mesgar ,K Swarnkar, E Simpson, and I Gurevych. 2019. Text processing like humans do: Visually attacking and shielding NLP systems[c]. *NAACL-HLT* 1634–47.

Gao, Y., G. Liang, F. Jiang, C Xu, J Yang, JR Chen, and H Wang. 2020. Social network rumor detection: A survey[J]. *ACTA ELECTONICA SINICA* 48 (7):1421.

Goodfellow, I. J., J. Shlens, and C. Szegedy. 2015. Explaining and harnessing adversarial examples[c]. *ICLR* 1–11.

Gupta, A. K., V. Paliwal, A. Rastogi, and P. Gupta. 2022. TRIESTE: Translation based defense for text classifiers[j]. *Journal of Ambient Intelligence and Humanized Computing* 1–12. doi:10.1007/s12652-022-03859-0.

He, X., L. Lyu, Q. Xu, and L Sun. 2021. Model extraction and adversarial transferability, your Bert is vulnerable![c]. *NAACL-HLT* 2006–12.

Hládek, D., J. Staš, and M. Pleva. 2020. Survey of automatic spelling correction[j]. *Electronics* 9 (10):1670. doi:10.3390/electronics9101670.

Jin, D., Z. Jin, Z. J. T, and P. Szolovits. 2020. Is Bert really robust? A strong baseline for natural language attack on text classification and entailment. *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (5):8018–25. doi:10.1609/aaai.v34i05.6311.

Li, J., S. Ji, T. Du, and T Wang. 2019. Textbugger: Generating adversarial text against real-world applications[c]. *NDSS* 1–15.

Ling, X., S. Ji, J. Zou, J Wang, C Wu, B Li, and T Wang. Deepsec: A uniform platform for security analysis of deep learning model[c] 2019 IEEE Symposium on Security and Privacy (SP). IEEE, 2019: 673–90.

Li, L., D. Song, J. Zeng, R Ma, and X Qiu. 2022. Rebuild and ensemble: Exploring defense against text adversaries[J]. *arXiv preprint arXiv:220314207*.

Madry, A., A. Makelov, L. Schmidt, D Tsipras, and A Vladu. 2018. Towards deep learning models resistant to adversarial attacks[j]. *ICLR* 1–18.

Ma, J., W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha. 2016. Detecting rumors from microblogs with recurrent neural networks[c]. *IJCAI* 3818–24.

Miyato, T., A. M. Dai, and I. Goodfellow. 2017. Adversarial training methods for semi-supervised text classification[c]. *ICLR* 1–12.

Ren, S., Y. Deng, K. He, and W Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. *Proceedings of the 57th annual meeting of the associationc for computational linguistics* 1085–97.

Ruchansky, N., S. Seo, and Y. Liu. 2017. Csi: A hybrid deep model for fake news detection[c]. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* 797–806.

Shu, K., L. Cui, S. Wang, et al. Defend: Explainable fake news detection[c] Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019: 395–405.

Shu, K., A. Sliva, S. Wang, J. Tang, and H. Liu. 2017. Fake news detection on social media: A data mining perspective[j]. *ACM SIGKDD Explorations Newsletter* 19 (1):22–36. doi:10.1145/3137597.3137600.

Si, C., Z. Zhang, F. Qi, et al. 2020. Better robustness by more coverage: Adversarial training with mixup augmentation for robust fine-tuning[j]. *arXiv preprint arXiv:2012 15699*.

Wang, Y., and M. Bansal. 2018. Robust machine comprehension models via adversarial training[j]. *arXiv preprint arXiv:1804 06473*.

Wang, W., P. Tang, J. Lou, et al. 2021. Certified robustness to word substitution attack with differential privacy[c].*Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 1102–12.

Wang, W. Y., and D. Yang. 2015. That's so annoying!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets[c]. *Proceedings of the 2015 conference on empirical methods in natural language processing* 2557–63.

Ye, M., C. Gong, and Q. Liu. 2020. SAFER: A structure-free approach for certified robustness to adversarial word substitutions[c]. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* 3465–75.

Yuan, C., Q. Ma, W. Zhou, et al. Jointly embedding the local and global relations of heterogeneous graph for rumor detection. 2019 IEEE international conference on data mining (ICDM). IEEE, 2019: 796–805.

Zang, Y., F. Qi, C. Yang, et al. 2019. Word-level textual adversarial attacking as combinatorial optimization[j]. *arXiv preprint arXiv:1910 12196.*

Zhou, Z., H. Guan, M. M. Bhat, et al. 2019. Fake news detection via NLP is vulnerable to adversarial attacks[j]. *arXiv preprint arXiv:1901 09657.*

Zhu, C., Y. Cheng, Z. Gan, et al. 2019. Freelb: Enhanced adversarial training for natural language understanding[j]. *arXiv preprint arXiv:1909 11764.*