Weibo Disaster Rumor Recognition Method Based on Adversarial Training and Stacked Structure

Lei Diao[1], Zhan Tang[1], Xuchao Guo[1], Zhao Bai[1], Shuhan Lu[2] and Lin Li[1*]
[1] College of Information and Electrical Engineering, China Agricultural University, Beijing,China
[e-mail: S20193081368@cau.edu.cn, tz_blues@163.com, gxc@cau.edu.cn, S20193081367@cau.edu.cn, lilincau@126.com]
[2] School of Information, University of Michigan
Ann Arbor, MI 48109, USA
[e-mail: shuhanlu@umich.edu]
*Corresponding author: Lin Li

Received February 9, 2022; revised August 14, 2022; accepted October 5, 2022; published October 31, 2022

Abstract

To solve the problems existing in the process of Weibo disaster rumor recognition, such as lack of corpus, poor text standardization, difficult to learn semantic information, and simple semantic features of disaster rumor text, this paper takes Sina Weibo as the data source, constructs a dataset for Weibo disaster rumor recognition, and proposes a deep learning model BERT_AT_Stacked LSTM for Weibo disaster rumor recognition. First, add adversarial disturbance to the embedding vector of each word to generate adversarial samples to enhance the features of rumor text, and carry out adversarial training to solve the problem that the text features of disaster rumors are relatively single. Second, the BERT part obtains the word-level semantic information of each Weibo text and generates a hidden vector containing sentence-level feature information. Finally, the hidden complex semantic information of poorly-regulated Weibo texts is learned using a Stacked Long Short-Term Memory (Stacked LSTM) structure. The experimental results show that, compared with other comparative models, the model in this paper has more advantages in recognizing disaster rumors on Weibo, with an F1_Socre of 97.48%, and has been tested on an open general domain dataset, with an F1_Score of 94.59%, indicating that the model has better generalization.

Keywords: Weibo disaster, Rumor recognition, Adversarial training, BERT, Stacked LSTM

This work was supported by the Major S&T project (Innovation 2030) of China under grant 2021ZD0113702

http://doi.org/10.3837/tiis.2022.10.001 **ISSN : 1976-7277**

1. Introduction

Disaster is suffering caused by natural or man-made scourge[1]. China is one of the countries with the most serious natural disasters in the world. There are many types of disasters, high frequency, wide distribution, and large losses. According to statistics, from 1999 to 2012, the average number of people affected by various natural disasters in China exceeded 300 million people each year, more than 3 million houses collapsed, more than 10 million people were urgently transferred and resettled, and the direct economic losses due to disasters exceeded 2,000. 100 million yuan[2], which has had a serious impact on the production and life of the people and the operation of the national economy. Among them, floods and earthquakes are the two most serious disaster types. In 1998, the flood in the whole basin of China caused disasters to 186 million people and killed 4,150 people[3]. In 2018, catastrophic flooding in Japan killed 220 people[4]. The floods that broke out in southern China in 2020 affected 30.2 million people and 141 people were killed or missing. On April 20, 2013, the Sichuan Lushan earthquake with a magnitude of 7.0 killed 196 people, left 21 missings, and injured 13,484 people. A total of 2.31 million people were affected[5]. As the losses caused by such disasters are increasing, the public is paying more and more attention to the real information of such disasters. The former social media tool has become a ubiquitous data warehouse with massive heterogeneous and real-time changes[6]. In China, Weibo is the main social media platform. When floods and earthquakes occur, a large amount of disaster information will be released and disseminated through this platform. The wide application of Weibo not only facilitates the public to release information and communicate with each other; but also provides a more convenient way for the release and dissemination of rumors. According to the data released by the official Weibo rumor refutation which is the official account of Weibo platform dedicated to refuting rumors, a total of 77,742 pieces of false information were effectively handled by Weibo in 2019, and rumors about disaster events are also common.

Disaster events have the characteristics of suddenness, destructiveness, uncertainty, urgency, and insufficient information[7], and the disaster event itself has two elements of rumor dissemination: importance and ambiguity[8]. Therefore, when a disaster occurs, not only various Weibo describing the real disaster situation will appear on Weibo, but also some false disaster rumors will be published and circulated. Disaster-related rumors on Weibo can be broadly classified into three categories: rumors based on fear and anxiety about death; rumors of "prey" that question real events; rumors rooted in historical religions and superstitions[9]. These disaster rumors flooded in Weibo are the most serious in the rumor crisis. They endanger the psychology and feelings of the public and also cause secondary harm to the people in disaster-stricken areas. More importantly, they provide support for subsequent disaster relief and disaster work misleading information, which greatly delayed the post-disaster rescue work. Therefore, recognizing disaster rumors in Weibo is of great significance for social stability and disaster relief.

For rumor recognition on social platforms, many scholars have carried out a lot of research, mainly based on machine learning and deep learning to carry out rumor recognition model research. Traditional machine learning methods mainly include Bayesian, decision tree, SVM, and so on. Although certain results have been achieved, it still relies on a large amount of feature engineering, which is time-consuming and labor-intensive. In recent years, deep learning methods have been widely used in the field of NLP, which avoids manual feature extraction and can make full use of the semantic information of the text itself. Some deep learning models such as CNN, LSTM, GRU, etc. are also used for rumor recognition in social networks. However, in the field of disaster rumor recognition, there are still the following

problems in rumor recognition:

(1) There are few public datasets of rumors available in the disaster area, which limits the research on the recognition of disaster rumors in the disaster area.
(2) The Weibo texts posted by individuals vary in length and are poorly standardized. It is generally difficult to systematically learn their effective semantic features by using word segmentation methods to obtain word vectors.
(3) Disaster rumors have few texts and single semantic features, which pose a great challenge to the generalization of the model.
(4) The traditional method of rumor feature extraction is relatively simple, mostly by artificially extracting social features and semantic features to expand information, but there is no further research on the semantic features of Weibo text itself. It is unable to effectively use a single deep learning model to deeply learn the semantic features of rumor texts to complete rumor recognition.

To solve the above problems, this paper constructed and annotated a Weibo flood-earthquake disaster rumor dataset to solve the problem of dataset scarcity. A BERT_AT_StackedLSTM model based on adversarial training[10] and stacked LSTM structure[11] was proposed. The BERT part of the model learns the word-level semantic information of the Weibo text, and the internal Transformer can better complete the feature extraction task. Adversarial training can solve the problem of a single feature with fewer rumor data, and data augmentation can effectively enhance the generalization of the model by adding disturbances to carry out adversarial training. The Stacked LSTM layer can better learn the hidden semantic information of irregular Weibo texts and improve the classification performance of the model.

The main work and contributions of this paper are as follows:

(1) The Weibo flood-earthquake rumor dataset is constructed for disaster rumor recognition task, which provides a new data foundation for research in this field.
(2) This paper proposes a deep learning model BERT_AT_Stacked LSTM, which takes the BERT model as the baseline model and adds adversarial perturbations to the embedding layer for adversarial training. These two mechanisms solve the problem that the semantic features are lost due to word segmentation errors and the semantic features are relatively single and prone to overfitting due to fewer rumors.
(3) Use a stacked LSTM structure for temporal and spatial modeling of character expression vectors. It solves the problems of poor standardization of social media texts and difficult extraction of semantic features.

The rest of this article is organized as follows: Section 2 is an introduction to related work that introduces the current research basis. Section 3 introduces the dataset constructed in this paper, and Section 4 introduces the Weibo disaster rumor recognition model. The experimental results and analysis are presented in Section 5. Section 6 summarizes this paper and introduces future work.

2. Related works

The methods used in early Weibo rumor recognition were mostly traditional machine learning methods. Wang et al. combined the @, #, URL, and other features in the Twitter text with the user's social attribute features (the number of friends, the number of fans, the user's reputation), and used a Bayesian classification model to recognize spam in Twitter[12]. Castillo et al. extracted the text features in Twitter for the first time, combined user feature topic features and propagation features, and used the J48 decision tree to complete the rumor recognition

task of Twitter text[13]. Qazvinian et al. used Twitter Nonitor tool for regular keyword matching to obtain tweets related to rumors, analyzed users' trust in rumors, and used Bayesian classifiers to identify rumors by combining text features, network features and propagation features[14]. By tracking the changes in the number of retweets of a rumor in different time periods, Tetsuro et al. summarized a series of change characteristics, such as whether there was a sudden surge, retweet rate, word distribution characteristics and text characteristics, etc., to achieve the identification of rumors[15]. Ziming Zeng et al. adopted the LDA topic model to obtain the topic features of Weibo texts, combined user credibility and Weibo influence features and proposed a rumor recognition model based on a random forest algorithm[16]. Yang et al. first proposed two features of publishing information client type and event location and combined the existing features to train a Weibo rumor classification model through SVM[17]. Based on Yang, Gang He et al. proposed four features: symbol feature, link feature, keyword distribution feature, and time difference[18]. Li et al. used a naive Bayesian model as a classifier to combine text features with the similarity between users to identify Weibo rumors[19]. This type of method mainly relies on manually extracted features. Although it has achieved good results, feature engineering takes a lot of time, and the increasing features are no longer effective in improving recognition accuracy.

In recent years, as deep learning[20] methods have been extensively studied in various fields, more scholars have begun to try to use deep learning methods to achieve the task of rumor recognition. Ma et al. first proposed to model Weibo text in time series and constructed a Weibo rumor recognition model based on RNN[21]. Zhiyuan Liu et al. designed a rumor early detection model using CNN[22]. Wenjing Ren et al. used LSTM and GRU to deal with the task of Weibo rumor recognition and achieved good results[23]. Mengjun Gao proposed a Weibo rumor recognition model based on attention mechanism[24] and LSTM to solve the problem of medium and long text dependence in Weibo rumor recognition[25]. Ao Li et al. first proposed a rumor detection model based on a generative adversarial network, which strengthens the learning of text features through the mutual confrontation between the generator and the recognizer[26]. Ma et al. constructed a top-down tree structure model to simulate the spread of Weibo rumors, combined with recurrent neural network (RvNN) for rumor recognition and classification[27]. In addition, more current research is to perform feature fusion between artificially extracted features and text features obtained by using deep learning models to complete rumor recognition. Lizhao Li et al. vectorized Weibo comment information obtained feature information through CNN, and then combined text features to realize rumor recognition[28]. Xuejian Huang et al. used the fusion of user features and text features and proposed a text feature extraction model based on Bi-GRU with attention mechanism[29]. Ran Sun et al. obtained the user characteristics, time characteristics, microblog text structure characteristics, text semantic characteristics and microblog propagation characteristics of rumored microblogs, and combined the BERT+CNN/RNN model to obtain textual features to improve the recognition rate of the rumor recognition model[30]. Such methods rely too much on artificially acquired features, despise the hidden information of the rumor text itself and the related characteristics of Weibo text, and use the deep learning model for extracting text features without great innovation. And, failed to combine the characteristics of rumor texts to propose a deep learning model to deal with the task of rumor recognition alone.

In addition, in the field of disasters, the task of recognizing rumors mainly relies on the research on rumor debunking in the field of sociology[8]. The main body of refuting rumors includes the government, the media, experts, parties, etc[31]. Puneet et al. proposed a game-theoretic approach to minimize the social impact of rumors of disaster events on the Internet[32]. Kyle et al. used several traditional machine learning algorithms to rely on hand-extracted tweet and publisher features to identify rumors in disaster events in Twitter[33]. The work is relatively heavy and the process is relatively complicated, and there are few studies on automatic recognition combined with computer-related technologies. Therefore, this paper proposes a disaster rumor recognition model based on adversarial training and stacked LSTM structure according to the characteristics of disaster information based on the existing research foundation of Weibo rumor recognition.

3. Dataset

3.1 Corpus collection

Disaster Weibo is mainly divided into real disaster Weibo and false disaster Weibo, the latter being disaster rumor Weibo. Nowadays, due to the huge number of Weibo users, a large number of new blog posts are published every day, and these Weibo posts are mixed, with good and bad news. The Weibo Community Convention has been established, and a Weibo Community Management Center has been established, where reported Weibo posts are processed and the results are made public. The Weibo information published by the Weibo Community Management Center is all false information that has been manually reviewed. For example, “Changfengyuan Mingdli Door-to-door euthanasia for pets” has been reviewed as a rumor text, and this text information is also the main data source of disaster rumors in this paper. For real disaster information, relevant Weibo information is obtained by subject keyword search through the Weibo homepage.

This paper use Sina Weibo's Weibo Community Management Center and Sina Weibo webpage as the corpus and use subject words to crawl more than 30,000 Weibo rumors and related real disaster situations related to floods and earthquakes in the past 10 years. Finally, 2,000 Weibo are obtained as the basic corpus through manual screening and other operations, and a dataset of Weibo flood-earthquake disaster rumors is constructed.

3.2 Dataset Labeling and Partitioning

The obtained 2000 Weibo texts were manually annotated, using the traditional two-category labeling method, with "0" and "1" as the labels of the disaster rumor Weibo and the real disaster Weibo. Among them, there were 800 Weibo posts about the false disaster situation and 1,200 Weibo posts about the real disaster situation. According to the ratio of 8:1:1, this paper selected 1600 Weibo in the dataset as the training set, 200 Weibo as the validation set, and 200 Weibo as the test set. The distribution of real information to rumor information in various datasets is shown in **Fig. 1**.

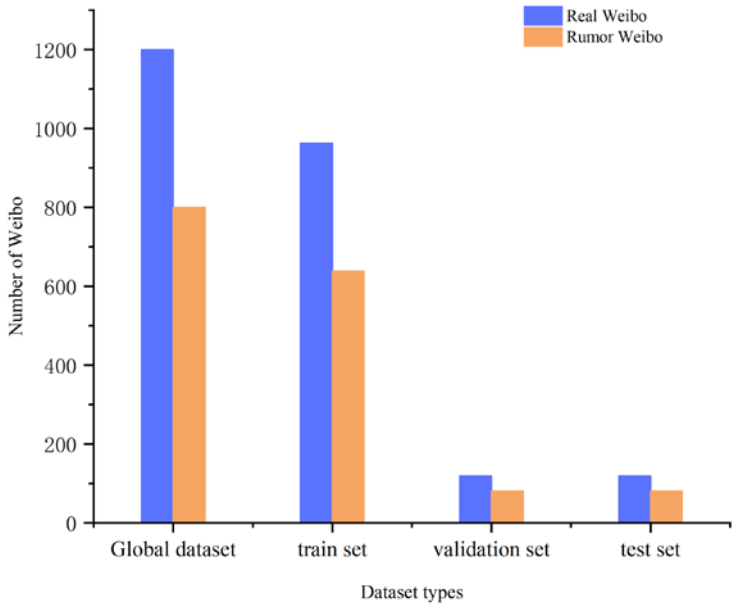


Fig. 1. Distribution map of real Weibo and rumored Weibo in each dataset

3.2 Dataset Features

Compared with the general field Weibo rumor dataset (the Weibo rumor dataset constructed by Zhiyuan Liu et al.[22]), the dataset of Weibo flood-earthquake rumors constructed in this paper has great differences in terms of text content field and professionalism. It has the following characteristics:

(1) It is more characteristic of disaster field. Different from the general field Weibo rumor dataset that contains rumors and real Weibo such as life, medicine, disaster, and entertainment, the Weibo flood-earthquake rumor dataset is all blog posts in the disaster field, and the boundaries between Weibo texts are relatively blurred.

(2) It has many kinds of disaster rumors. The rumor dataset includes exaggerated disaster situations, new use of old news, disaster relief rumors, secondary disaster rumors, irrelevant information (Weibo that contains disaster keywords when a disaster occurs but have nothing to do with the disaster), etc.

(3) Rumor data accounts for less. Different from the public dataset, the ratio of rumor Weibo to real Weibo is close to 1:1, the Weibo flood-earthquake rumors dataset only accounts for 2/5 of the total number of rumors, and the semantic features are relatively single, which increases the difficulty of learning semantic features for the model. And, it also poses a challenge to the generalization of the model.

Table 1 is an example of relevant data.

Table 1. Some data examples of the dataset

Disaster Category	Label	Rumors or Real rmation	Weibo Text
Flood	0	Rumor	江西洪水决堤， 死了太多人了！！！ (The floods in Jiangxi burst, and too many people died! ! !)
Flood	1	Real	广州的洪涝也有 点严重了吧?原图? (The floods in Guangzhou are also a bit serious, right? Original picture?)
Earthquake	0	Rumor	四川卫视:一位叫徐敬的 女孩，21 岁，请速回雅安 水城县人民医院，妈妈伤的 很严重，想见你最后一面，爱 心接力，请即转发’祝愿平安 (Sichuan Satellite TV: A girl named Xu Jing, 21 years old, please return to Ya'an Shuicheng County People's Hospital as soon as possible, my mother is seriously injured, I want to see you for the last time, love relay, please forward 'wish peace)
Earthquake	1	Real	为四川阿坝地震灾区祈福，愿灾 难远去，少一些伤亡，多一些平 安！#四川省地震# |张家港 (Pray for the earthquake-stricken area in Aba, Sichuan. May the disaster go away, with fewer casualties and more peace! #Sichuan earthquake# | Zhangjiagang)

4. Method

Aiming at the problems existing in the recognition task of Weibo disaster rumors, this paper takes each word in Weibo disaster text as input and proposes a deep learning model BERT_AT_Stacked LSTM suitable for Weibo disaster rumor recognition. The main structure of the model is shown in **Fig. 2**, which mainly includes the addition of the adversarial training part, the BERT part, and the Stacked LSTM part. As shown in **Fig. 2**, the model input is a disaster Weibo text: “#四川7. 0级地震#愿生者平安坚强，愿死者安息。（#Sichuan7.0 earthquake# May the living be safe and strong, and the dead may rest in peace.）”. After word splitting and padding and adding adversarial perturbation, n initial expression vectors W_{ri}（1≤i≤n）are obtained. Then, after the BERT processing, the expression vector T_i of each word is obtained, and the final recognition result is obtained after the Stacked LSTM layer calculation processing and adversarial training.

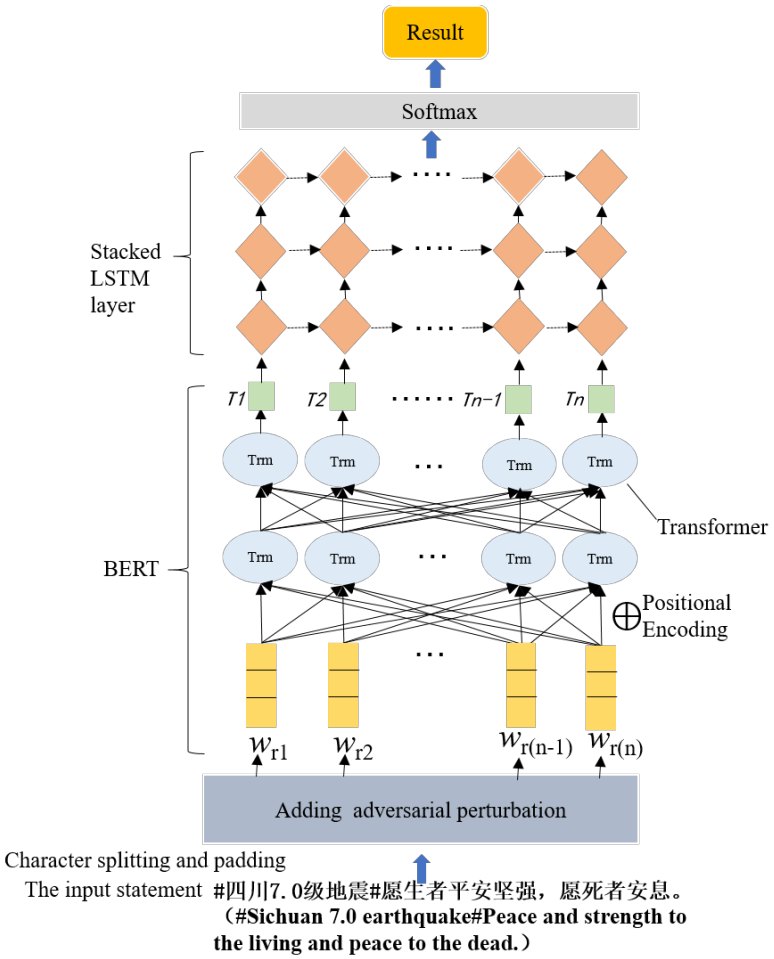


Fig. 2. Model structure diagram

4.1 Adversarial training

To take full advantage of the word-level semantic information of Weibo texts, this paper uses Google's Chinese BERT pre-training model to fine-tune to generate hidden vectors of individual characters in each Weibo text. Each sentence of Weibo text processed by the BERT word embedding layer can be represented as $B = [W_1, W_2, \ldots, W_i, W_n]$, $W \in R^{d\times 1}$. Among them, W_1 and W_n represent special characters [CLS] and [SEP], which represent categorical features and clause symbols, respectively. In addition to these two special values, W_i ($i \in [2, n-1]$) represents the initial word embedding vector of the i-th word in the disaster Weibo text.

To solve the problem of fewer disaster rumors, single semantic features, and easy overfitting, this paper adds adversarial training to the model, and actively adds noise to the embedding layer to generate new adversarial samples as a data enhancement method. As shown in **Fig. 3**, the adversarial perturbation within the adversarial range is added to the initial character vector after the embedding layer, as shown in **Fig. 3** r, $||r|| \leq \epsilon$, where ϵ is a bounded constant. That destroys the original semantic information and generates new interference data, thereby enhancing the recognition performance of the model. The initial word embedding vector W_i of each word in B becomes $B_r = [W_{r1}, W_{r2}, \ldots, W_{ri}, W_{rn}]$ after

adding adversarial perturbation, and its calculation formula is:

$$W_r = W + r \tag{1}$$

Adversarial training, essentially a new regularization method, aims to increase model robustness and generalization by adding approximate worst perturbations. This paper takes W as input and θas model parameter. When applied to a model, adversarial training adds the following formula to the cost function.

$$-\log p(y|W + r_{adv}; \theta) \tag{2}$$

$$r_{adv} = \arg\min \log p(y|W + r\ ;\hat{\theta}) \tag{3}$$

The r is an input disturbance, θ is the model weight, $\hat{\theta}$ is the current weight of the model, and y is the predicted label value. The training objective is to minimize the loss function value by training θ. During the model training process, we determine the approximate worst perturbation r_{adv} for the current model through $p(y|W + r\ ;\hat{\theta})$ in (3) and update the parameter θ in the model to obtain the minimum value through training to make the model robust to this perturbation, thus accomplishing our training purpose. Since (3) is no-differentiability, we adopt the following formula to approximate r_{adv}.

$$r_{adv} = -\frac{\epsilon g}{||g||_2} \tag{4}$$

$$g = \nabla_w \log p(y|W; \hat{\theta}) \tag{5}$$

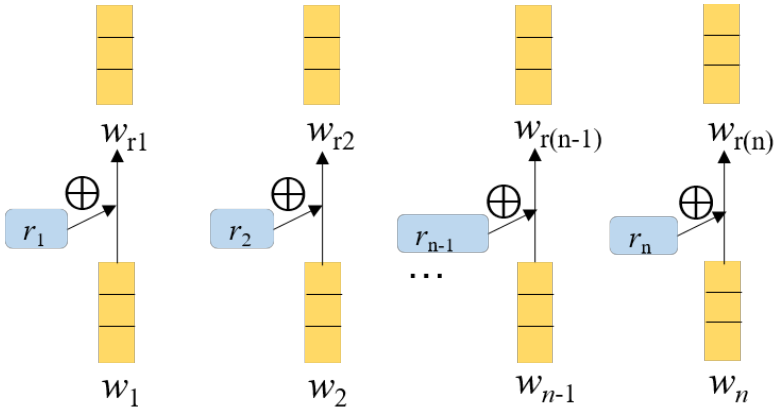


Fig. 3. Initial word vector adding adversarial perturbation

4.2 BRRT

Denote a set of encoding vectors containing the position information of each character in the Weibo text obtained by B_r after position encoding as $E = [e_1, e_2, \ldots, e_i, e_{n-1}, e_n]$, $e_i \in R^{d\times 1}$. Input E into a multi-layer self-attention-based[34] Transformer whose structure is shown in **Fig. 4**. The most important part of the Transformer is the multi-head self-attention mechanism, which is mainly obtained by adjusting the weight coefficient matrix by the degree of association between words in the same sentence. The operation formula of the self-attention mechanism is as follows:

$$Q = \text{Linear}(E) = EW_Q \tag{6}$$

$$K = \text{Linear}(E) = EW_k \tag{7}$$

$$V = \text{Linear}(E) = EW_V \tag{8}$$

Q、K、V—the linear mapping matrix of E, W_Q, W_k, W_V—the assigned weight matrix, E is processed by the multi-head self-attention mechanism to obtain the self-attention weight. After E is processed by the multi-head self-attention mechanism, the self-attention weight is obtained, and a new character representation vector $E_{attention}$ is obtained through residual linking and normalization.

$$E_{attention} = Selfattention(Q, K, V) \tag{9}$$

$$E_{attention} = E + E_{attention} \tag{10}$$

$$E_{attention} = LayerNorm(E_{attention}) \tag{11}$$

Representation vector $E_{attention}$, and then through the feedforward network layer and residual link and normalization to obtain the final character hidden vector $T = [T_1, T_2, \ldots T_{n-1}, T_n]$, $T_i \in R^{b*l*d}$ where b is the batch_size size and l is the sentence length. The operation process is shown in (12) (13) (14):

$$E_{hidden} = Activate(Linear(Linear(E_{attention}))) \tag{12}$$

$$E_{hidden} = E_{hidden} + E_{attention} \tag{13}$$

$$T = LayerNorm(E_{hidden}) \tag{14}$$

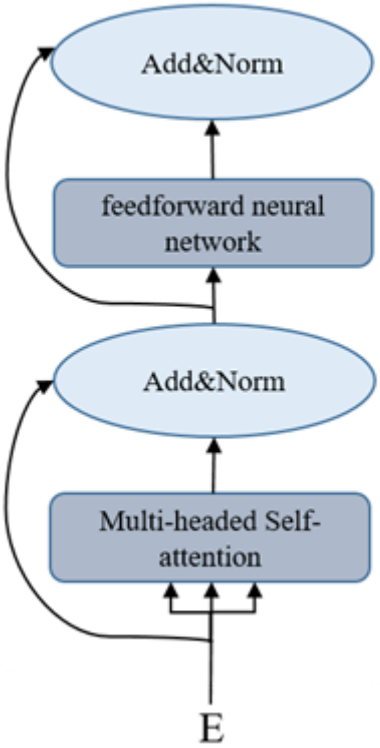


Fig. 4. Transformer structure diagram

4.3 Stacked LSTM layer

This layer is mainly used to extract the hidden semantic information of the disaster Weibo text, and solve the problems of poor standardization of Weibo text and difficulty to extract semantic information. To determine whether a disaster Weibo is a rumor, this paper inputs the word

embedding expression vector T generated by BERT into the LSTM network layer of the stacked structure. Unlike the traditional single-layer LSTM that only accepts the current word embedding expression vector and hidden state vector, the input of Stacked LSTM becomes the hidden state vector of the LSTM output of the previous layer and the hidden state vector of the same layer at the previous moment from the second layer. This stacked structure makes up for the shortcomings of the single-layer LSTM spatial modeling ability. Through deeper learning, enhances the model's ability to learn text semantic information, and solves the problem of sparse features, semantic information is difficult to obtain in poorly standardized text and lack of spatial modeling ability has effectively improved the performance of the model to identify disaster rumors on Weibo.

Let the representation vector of the i-th word in the disaster Weibo text be T_i, and $\acute{h}_{j,i}$ represents the output hidden layer vector of the j-th layer in the Stacked LSTM layer at the time i. The formula for generating hidden layers of Stacked LSTM can be expressed as:

First layer:

$$\acute{h}_i = LSTM(T_i, h_{i-1}) \tag{15}$$

The j-th layer (j >1):

$$\acute{h}_{j,i} = LSTM(h_{j,i-1}, h_{j-1,i}) \tag{16}$$

Finally, the character expression hidden vector H output by the last LSTM unit of the last layer is processed by the Softmax function to obtain the final recognition probability. The formula is as follows:

$$\mathrm{p} = \mathrm{Softmax(DH)} \tag{17}$$

5. Experiment

5.1 Experiment setup

This paper used Google's Chinese BERT pre-training model Chinese_L-12_H-768_A-12 for fine-tuning to generate the hidden vector of a single word in each Weibo disaster text. The Batch_Size was set to 4, and the size of Adam was equal to 2×10^{-5} to minimize the training target. The cross-entropy was used as the loss function. The model in this paper adopted the Keras2.3.1 framework and the operating environment was GTX1660. To better evaluate the performance of the model, this paper selected the following six deep learning models to compare with the new model proposed in this paper.

- CNN: As a common model in the image field, CNN has also achieved good results in text classification in recent years, and has also shown good results in Weibo rumor recognition.
- LSTM: This model is an optimization of RNN, which learns the semantic information of text through time series modeling, and is widely used in rumor recognition tasks.
- GRU: This model is an optimization of the LSTM network and is widely used in a variety of text classification tasks, including the second classification of rumor text and real text.
- Attention_LSTM: The model is a combination of LSTM and attention mechanism. And Zhiwei Jin et al. [25] used this network to deal with the task of Weibo rumor recognition.
- BERT: Based on the pre-trained language model of Transformer, the first-word embedding expression vector generated by each disaster Weibo text is selected as the global information, to complete the rumor recognition.
- BERT_Stacked LSTM[35]: An improved model of the BERT model, commonly used to perform various complex text classification tasks.

5.2 Evaluation Metrics

In this paper, the precision P, the recall rate R, and the F1_Score are used to evaluate the model performance, which is defined as follows:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{18}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{19}$$

$$\text{F1} = \frac{2TP}{2TP+FP+FN} \tag{20}$$

Among them, TP is the number of predicted positive classes as positive classes; FP is the number of predicted negative classes as positive classes; FN is the number of predicted positive classes as negative classes.

5.3 Experimental results and analysis

5.3.1 Results on Weibo flood-earthquake Rumors Test Set

This section report the experimental results of the BERT_AT_Stacked LSTM model on the Weibo flood-earthquake rumor dataset.

Table 2. Experimental results on the test set

Model	P	R	F1
CNN	95.77	94.02	94.72
LSTM	93.61	91.94	92.61
GRU	91.06	89.26	89.94
Attention_LSTM	93.78	93.78	93.78
BERT	93.55	97.48	95.47
BERT_Stacked LSTM	94.35	98.32	96.30
BERT_AT_Stacked LSTM	**97.48**	**97.48**	**97.48**

As shown in **Table 2**, it can be seen that compared with the LSTM based on time series modeling which F1_Score is 92.61%, the CNN with stronger spatial modeling ability has a better recognition effect, and the F1_Score is 94.72%, which is 2.11 percentage points higher than that of LSTM. It shows that in the task of recognizing disaster rumors on Weibo, the ability of spatial modeling plays a better role in the recognition effect. When using the GRU model, the F1_Score decreased by 2.67 percentage points compared with LSTM to 89.94%, indicating that the GRU, which performed better in processing simple text binary classification, did not achieve better results in the task of recognizing Weibo disaster rumors. When the attention mechanism is added based on LSTM, the F1_Score is 93.78%, which is 1.17 percentage points higher than that of LSTM, indicating that after introducing the attention weight, the problem of long text dependence is solved and the recognition performance of the model is improved.

When the pre-trained language model BERT is used, the recognition effect is significantly improved, and the F1_Score is 95.47%, it of the first three comparison models is increased by 0.75 percentage points, 2.86 percentage points, and 5.53 percentage points respectively. It shows that after using the word-level embedding vector, the problem of missing semantics caused by word segmentation errors is alleviated to a certain extent, thereby improving the performance of rumor recognition. BERT_Stacked LSTM, which models the hidden vectors of the disaster Weibo text word embedding generated by BERT according to time series, and the unique stacking structure also increases the spatial modeling ability of the model, that

effectively solves the problem of poor standardization of Weibo text and difficult learning of semantic information, thereby increasing the F1_Score to 96.30%.

In addition, this paper adds adversarial training based on BERT_Stacked LSTM and adds adversarial disturbance to the embedding layer to generate adversarial samples, which solves the problem that the rumor text in the disaster area is relatively simple, and the features are less likely to cause the model to overfit and have poor robustness, that obtains the optimal model BERT_AT_Stacked LSTM. The F1_Score of BERT_AT_Stacked LSTM on the test set is 97.48%, which is 1.18 percentage points higher than that of BERT_Stacked LSTM and better than all other comparable models.

5.3.2 Analysis of True and False Information Recognition Effect

Fig. 5 shows the F1_Score of real disaster information and disaster rumor information in each recognition model. As shown in **Fig. 5**, all models have a better ability to recognize real disaster information than to identify disaster rumors. Among all the basic models (CNN, LSTM, GRU, BERT, the F1_Score of real disasters is 2.01, 3, 4.03, 2 percentage points higher than that of disaster rumors, respectively), BERT has the best ability to recognize real disasters and disaster rumors, indicating that word-level learning of semantic information is more advantageous than word-segmentation learning. After using the stacking structure, the model mainly enhances the recognition ability of real disaster information, and its F1_Score is increased by 0.98 percentage points compared with BERT, but the recognition effect of disaster rumors is not significantly improved, and its F1_Score is only increased 0.62 percentage points, which shows that when the stacking structure learns more complex semantic information, and for disaster rumors with relatively simple features, it cannot improve its feature extraction ability. However, after adding adversarial training, the F1_Score of both real disaster information and disaster rumor information is significantly improved, reaching 98.22% and 96.74%, respectively, indicating that adversarial training helps the model extract features from real disaster information and disaster rumor information.

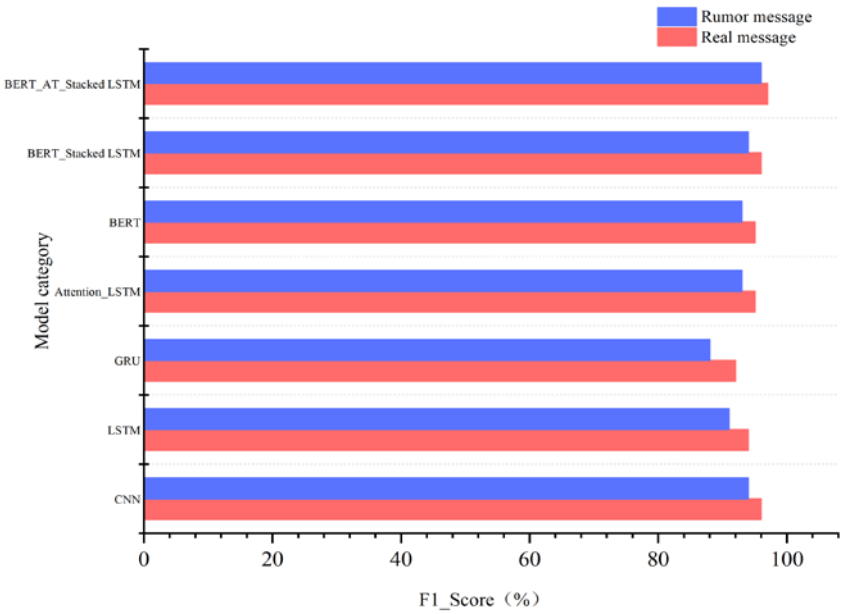


Fig. 5. Recognition renderings of real disaster information and disaster rumor information in various models

5.3.3 The effect of adversarial and stacked structures on recognition ability

In this paper, BERT is used as the baseline model for real and false disaster information recognition, and stacked LSTM and adversarial training are added based on BERT. The effect between the two and BERT on the recognition performance of the recognition model is shown in **Fig. 6**. The F1_Score of the BERT_AT model with adversarial training added to BERT is 0.33 percentage points higher than that of BERT alone, indicating that adversarial training improves the overall recognition ability of the model. Based on BERT, the BERT_Staceked LSTM model with stacked LSTM is added, which also effectively improves the recognition performance of the model, and its F1_Score is 0.83 percentage points higher than that of BERT alone, indicating the advantages of the stacked structure for non-normative texts. In addition, after combining the stacking structure and adversarial training, the recognition ability of the model has been optimally improved. BERT_AT_Stacked LSTM not only greatly outperforms the BERT model, but also significantly improves the performance compared to BERT_AT and BERT_Stacked LSTM, its F1_Score is 2.01, 1.68, and 1.18 percentage points higher than those of the three models, respectively.

.

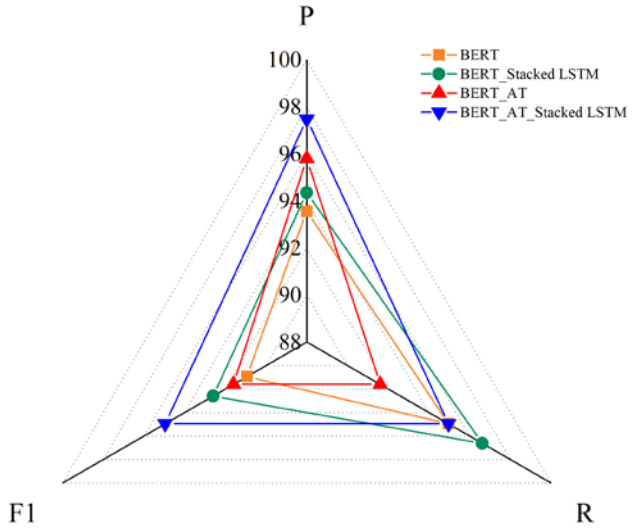


Fig. 6. Influence diagram of the stacked structure and adversarial training effect

5.3.4 The effect of the number of stacked layers on the recognition performance

One of the reasons for the excellent results of the Weibo disaster rumor recognition model proposed in this paper is that it adopts the stacked LSTM structure. This part analyzes the influence of the number of stacked layers on the performance of the model. In this paper, the number of stacked layers is gradually increased from 1 layer to 5 layers, and two models BERT_Stacked LSTM and BERT_AT_Stacked LSTM using the stacked structure are selected for comparative analysis. The relationship between the F1_Score and the number of stacked layers is shown in **Fig. 7**. As shown in **Fig. 7**, when the number of stacked layers is from 1 to 2, both the BERT_Stacked LSTM and BERT_AT_Stacked LSTM models show an upward trend in F1_Score, the changes are 93.02% to 96.3%, 95.35% to 97.05%, respectively,, and the BERT_Stacked LSTM reaches the optimal value at the second layer. When the number of layers is from 2 to 5 layers, the F1_Score of BERT_Stacked LSTM decreases significantly, the change in F1_Score is 96.3%→95.8%→92.44%→93% while BERT_AT_Stacked LSTM first

increases and then decreases, the change in F1_Score is 97.05%→97.48%→95.4%→95.44%, and reaches the optimal value at the third layer. This trend of increasing first and then decreasing is because when the stacked LSTM structure is used, the spatial modeling ability of the model is increased, and the problem that the semantic information of irregular and complex text is difficult to learn is solved, but when the number of layers reaches a certain value, the model will overfit and the effect will decrease. In addition, since adversarial training creates new adversarial samples by adding adversarial disturbances, which enriches relatively single Weibo disaster texts, the requirements for spatial capabilities are higher than that of BERT_Stacked LSTM, and the limit on the number of layers in overfitting will also increase, so BERT_Stacked LSTM appears to overfitting earlier than BERT_AT_Stacked LSTM. Therefore, this paper chooses 3 as the optimal number of layers for the BERT_AT_Stacked LSTM model.

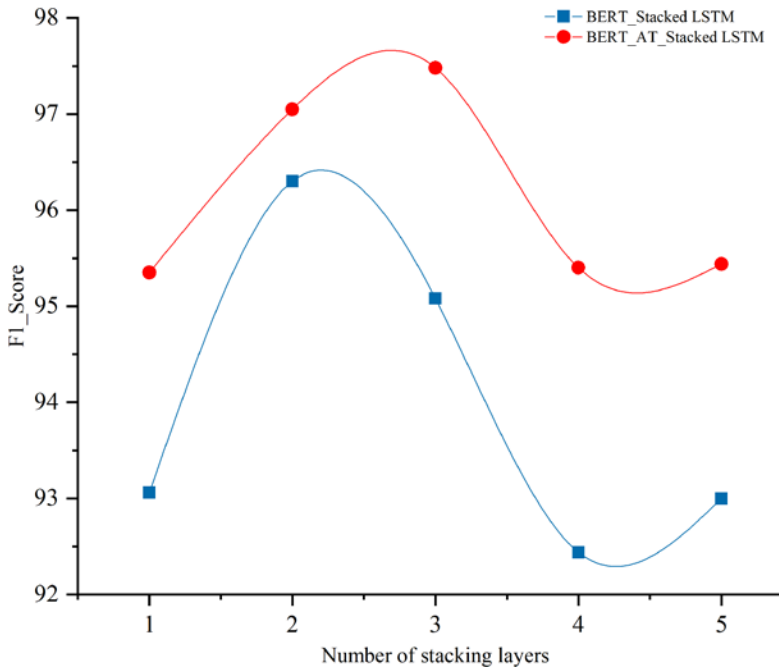


Fig. 7. The effect of stacking layers on the effect of BERT_Stacked LSTM and BERT_AT_Stacked LSTM model

5.3.5 Model generalization analysis

To verify the generalization of the model, this paper selects the Weibo rumor dataset constructed by Zhiyuan Liu of Tsinghua University[22], it is a general domain dataset, including various real and false Weibo, mainly including 1538 rumors and 1849 real Weibo texts. The test results on such a general-purpose dataset can fully demonstrate the excellent generalization of the model. The test results are shown in **Table 3**. It can be seen that the test results of BERT are significantly better than the previous deep learning models, with the F1_Score reaching 93.33%, while the F1_Score of the BERT_Stacked LSTM model is 0.33 percentage points higher than that of BERT. It is shown that the word-level pre-trained language model is suitable for use as a baseline model for rumor recognition problems, and the stacked structure can better learn the semantic information of Weibo texts. The optimal model of this paper, BERT_AT_Stacked LSTM, also performs the best when dealing with

rumor recognition tasks in the general domain, which proves that adversarial training can improve the overall generalization ability of the model.

Table 3. Public dataset test results

Model	P	R	F1	Real			Rumor		
				P	R	F1	P	R	F1
CNN	91.63	91.04	91.27	90.31	94.65	92.43	92.96	87.42	90.10
LSTM	85.73	84.92	85.19	84.50	90.37	87.34	86.96	79.47	83.05
GRU	88.18	88.84	87.98	88.08	90.91	89.47	88.27	84.77	86.49
Attention_LSTM	88.91	87.85	88.19	86.63	93.58	89.97	91.18	82.12	86.41
BERT	93.09	93.58	93.33	93.09	93.58	93.33	92.00	91.39	91.69
BERT_Stacked LSTM	94.54	92.51	93.51	94.54	92.51	93.51	90.97	93.38	92.16
BERT_AT_Stacked LSTM	95.63	93.58	94.59	95.62	93.58	94.60	92.26	94.70	93.46

6. Conclusion

To solve the problem of lack of the corpus in Weibo disaster rumor identification task, this paper constructs a Weibo flood-earthquake rumor dataset including 2000 disaster Weibo texts, including 1200 real disaster information and 800 disaster rumor information, and compared with general domain dataset, it is more professional and with more types of disaster information and fewer rumors. In addition, a new deep learning model BERT_AT_Stacked LSTM is extracted to solve the problems of poor text standardization, single rumor features, and excessive reliance on artificially extracted social attribute feature fusion in the Weibo disaster rumor recognition task. The F1_Score of the model reaches 97.48% on the self-constructed dataset, which greatly improves the recognition performance. And the generalization of the model is verified on public datasets.

In future work, we will focus on the following aspects:

(1) In view of the text characteristics of Weibo rumors in other fields, the deep learning model is continuously improved, and the features in specific fields are tried to be integrated so that it can complete the task of rumor recognition more accurately and efficiently.

(2) We will try to use the data enhancement feature of the adversarial network and combine the graph convolutional neural network to complete the fusion of rumor propagation features and the sufficient semantic features of the text itself to complete more complex and secret Weibo rumor recognition.

(3) We will try to build a comprehensive disaster rumor recognition dataset containing all kinds of disaster information, and build a new recognition model to solve the problem that the common features of multiple disaster texts are difficult to obtain.

References

[1] Z. H. Meng, C. R. Peng, *History of Famine in China*, Beijing, China: WREPP,1989.

[2] H. Li, "Research on Natural Disasters And Their Economic Costs in My Country," *Price monthly publication*, vol. 004, no. 000, pp. 66-72, 2010.

[3] Z. Q. Wan, "1998 Flood Disaster Report," *China Flood & Drought Management*, no. 04, pp. 04, 1998.

[4] R. Jang, "Japan's Great Flood in July 2018 and Its Response," *China Flood & Drought Management*, vol. 28, no. 8, pp. 09-12, 2018. Article (CrossRef Link)
[5] X. H. Su, X. D. Zhang, C. L. Hu, Z. C. Zou and X. K. Qiu, "Research on the Extraction of Earthquake's Hot Topic-Words form Microblog Based on Improved TF-IDF Algorithm", *Geography and Geo-Information Science*, vol. 34, no. 04, pp. 90-95, 2018.
[6] L. L. Ma, H. W. L, S. W. Lian, R. P Liang and J. Gong, "An Ontology Modeling Method for Natural Disaster Events," *Geography and Geo-Information Science*, vol. 32, no. 01, pp. 12-17, 2016.
[7] Z. C. Tian, "Research on Disaster News in Contemporary China," Ph.D. dissertation, Dept. Journalism, Fudan Univ., Shanghai, China, 2005. Article (CrossRef Link)
[8] X, Song, "Research on Multi-subject Rumor-refuting of Disaster Events Based on the Perspective of Collaborative Governance," Ph.D. dissertation, Dept. Journalism, ECNU Univ., Shanghai, China, 2020. Article (CrossRef Link)
[9] H. L. Cao, "Liminality and Rumor: A Religious Anthropological Interpretation of Earthquake Disaster," *Qinghai Social Sciences*, no. 03, pp. 131-135+117, 2010. Article (CrossRef Link)
[10] T. Miyato, A.M. Dai and I. Goodfellow, "Adversarial Training Methods for Semi-Supervised Text Classification," in *Proc. of ICLR*, 2017. Article (CrossRef Link)
[11] C. España-Bonet, J. A. R. Fonollosa, "Automatic Speech Recognition with Deep Neural Networks for Impaired Speech," in *Proc. of IberSPEECH 2016: Advances in Speech and Language Technologies for Iberian Languages*, pp. 97–107, 2016. Article (CrossRef Link)
[12] A. H. Wang, "Don't Follow Me - Spam Detection in Twitter," in *Proc. of SECRYPT 2010 - Proceedings of the International Conference on Security and Cryptography*, 2010. Article (CrossRef Link)
[13] C. Castillo, M. Mendoza and B. Poblete, "Information credibility on Twitter," in *Proc. of the 20th International Conference on World Wide Web*, pp. 675-684, 2011. Article (CrossRef Link)
[14] V. Qazvinian, E. Rosengren, D R Radev and Q. Mei, "Rumor has it: Identifying Misinformation in Microblogs," in *Proc. of EMNLP '11: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1589-1599, 2011. Article(CrossRef Link)
[15] T Takahashi, N. Igata, "Rumor detection on twitter," in *Proc. of The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems*, 2013. Article(CrossRef Link)
[16] Z. M. Zeng, J. Wang, "Research on Weibo Rumor Recognition Based on LDA and Random Forest——Taking 2016 Smog Rumors as an Example," *Journal of the China Society for Scientific and Technical Information*, vol. 38, no. 1, pp. 89-96, 2019. Article (CrossRef Link)
[17] F. Yang, X. Yu, Y. Liu, M Yang, "Automatic Detection of Rumor on Sina Weibo," in *Proc. of ACM SIGKDD*, pp. 1-7, 2012. Article (CrossRef Link)
[18] G. He, X. Q. Lv, Z. Li and L. P. Xu, "Research on Weibo Rumor Recognition," *Library and Information Service*, vol. 57, no. 23, pp. 114-120, 2013. Article (CrossRef Link)
[19] C. Li, F. Liu, P. Li, "Text Similarity Computation Model for Identifying Rumor Based on Bayesian Network in Microblog," *International Arab Journal of Information Technology*, vol. 17, no 5, pp. 731-741, 2020. Article(CrossRef Link)
[20] Y. Lecun, Y. Bengio, G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015. Article (CrossRef Link)
[21] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K. F. Wang and M. Cha, "Detecting Rumors from Microblogs with Recurrent Neural Networks," in *Proc. of International Joint Conference on Artificial Intelligence*, 2016. Article (CrossRef Link)
[22] Z. Liu, C. Song,C. Yang, "Early Automatic Detection of Rumors on Social Media Platforms," *Global Journal of Media Studies*, vol. 005, no. 004, pp. 65-80, 2018. Article (CrossRef Link)
[23] W. J. Ren, B. Qin, T. Liu, "Rumor Detection Based on Time Series Model," *Intelligent Computer and Applications*, vol. 09, no. 03, pp. 300-303, 2019. Article (CrossRef Link)

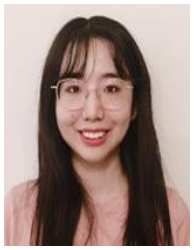
[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is All You Need," in *Proc. of NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017. Article (CrossRef Link)
[25] M. J. Gao, " Research on Weibo Rumor Recognition Based on Deep Learning," Ph.D. dissertation, Dept. Journalism, JUFE Univ., Changchun, China, 2021. Article (CrossRef Link)
[26] A. Li, Z. P. Dan,F. M. Dong, W. L. Liu and Y. Feng, "A Rumor Detection Method Based on Improved Generative Adversarial Network," *Journal of Chinese Information Processing*, vol. 34, no. 09, pp. 78-88, 2020. Article (CrossRef Link)
[27] J. Ma, W. Gao, S. Joty, K. Wong, "An Attention-based Rumor Detection Model with Tree-structured Recursive Neural Networks," *ACM Transactions on Intelligent Systems and Technology*, vol. 11, no. 4, pp. 1-28, 2020. Article(CrossRef Link)
[28] L. Z. Li, G. Cai, J. Pan, "A Microblog Rumor Events Detection Method Based on C-GRU," *Journal of Shandong University(Engineering Science)*, vol. 49, no. 02, pp. 102-106+115, 2019.
[29] X. J. Huang, G. S. Wang, Y. S. Luo, L Min, X. F. Wu and Z. P. Li, "A Real-time Detection Model of Weibo Rumors that Integrates Multi-user Features and Content Features," *Journal of Chinese Computer Systems*, pp. 1-12, 2021. Article (CrossRef Link)
[30] R. Sun, L. An, "Research on Rumor Recognition in Public Health Emergencies," *information work*, vol. 42, no.5, pp. 8, 2021. Article(CrossRef Link)
[31] Y. Xiong, "Set Up Scientific Procedures to Curb the Spread of Rumour of Emergencies," *News and Writing*, no. 07, pp. 16-19, 2012.
[32] P. Agarwal, R. A. Aziz, J. Zhuang, "Interplay of rumor propagation and clarification on social media during crisis events - A game-theoretic approach," *European Journal of Operational Research*, vol. 298, no. 2, pp. 714-733, 2022. Article(CrossRef Link)
[33] H. Kyle, P. Agarwal, J. Zhuang, "Monitoring Misinformation on Twitter During Crisis Events: A Machine Learning Approach," *Risk Analysis*, vol. 42, no, 8, pp. 1728-1748, 2022. Article(CrossRef Link)
[34] Z. H. Lin, M. W. Feng, C. N. Santos, M. Yu , B. Xiang, B. Zhou and Y. Bengio, "A Self-attentive Sentence Embedding," in *Proc. of ICLR 2017*, 2017. Article (CrossRef Link)
[35] L. Li, L. Diao, Z. Tang, Z. Bai, H. Zhou and X. C. Guo, "Question Classification Method of Agricultural Diseases and Pets Based on BERT_Stacked LSTM," *Transactions of the Chinese Society for Agricultural Machinery*, vol. 52, no. S1, pp. 172-177, 2021.

LEI DIAO received a bachelor's degree in engineering from Anhui University of Science and Technology. He is currently pursuing the M.S. degree with the College of Information and Electrical Engineering, China Agricultural University, Beijing, China. His research interests include natural language processing, text mining, and deep learning.

ZHAN TAN received the M.S. degree in computer science from Shanghai Normal University and his Bachelor's degree from Southwest Jiaotong University. He is currently pursuing the Ph.D. degree with the College of Information and Electrical Engineering, China Agricultural University, Beijing, China. His research interests include natural language processing, text mining, and deep learning.

XUCHAO GUO received his Bachelor's degree in Computer Science and his Master's degree in 2018 from Shandong Agricultural University. He is currently Ph.D. student in the College of Information and Electrical Engineering, China Agricultural University. He is mainly engaged in natural language processing and knowledge graph in agricultural fields. His research interests include: deep learning, complex network analysis, data mining, machine learning, and image processing.

ZHAO BAI received a bachelor's degree in engineering from Anhui University of Science and Technology. He is currently pursuing the M.S. degree with the College of Information and Electrical Engineering, China Agricultural University, Beijing, China. His research interests is computer vision.



SHUHAN LU received her Bachelor's degree in Computer and Information Science in 2020 at the Ohio State University. From 2020, she is currently studying Master's in Health Informatics in University of Michigan, Ann Arbor. She has good experience in database creating, database managing, and data analysis. Her research interests include: data mining, database manage, machine learning and image processing.

LIN LI is currently a Professor and a Doctoral Supervisor with the College of Information and Electrical Engineering (CIEE), China Agricultural University. Her main research interests include knowledge engineering and machine learning.