


# Emotion-Drive Interpretable Fake News Detection

Xiaoyi Ge, Engineering University of PAP, China

 <https://orcid.org/0000-0002-7252-7085>

Mingshu Zhang, Engineering University of PAP, China\*

Xu An Wang, Engineering University of PAP, China

Jia Liu, Engineering University of PAP, China

Bin Wei, Engineering University of PAP, China

## ABSTRACT

Fake news has brought significant challenges to the healthy development of social media. Although current fake news detection methods are advanced, many models directly utilize unselected user comments and do not consider the emotional connection between news content and user comments. The authors propose an emotion-driven explainable fake news detection model (EDI) to solve this problem. The model can select valuable user comments by using sentiment value, obtain the emotional correlation representation between news content and user comments by using collaborative annotation, and obtain the weighted representation of user comments by using the attention mechanism. Experimental results on Twitter and Weibo show that the detection model significantly outperforms the state-of-the-art models and provides reasonable interpretation.

## KEYWORDS

Co-Attention, Emotion Attention, Emotion Representation, Emotion Selection, Fake News Detection, Interpretable

## INTRODUCTION

While providing great convenience to people's daily life, social media also promotes the spread of fake news and has negative effects on society, economy, and culture. During major events such as the US presidential election (Allcott & Gentzkow, 2017), the COVID-19 pandemic (Diseases, 2020), and the Russian-Ukrainian conflict (Haq et al., 2022), social media platforms played an extremely critical role in distributing information while being bombarded with misinformation as well, such as a bunch of fake news. In this regard, the propagation of fake news must be detected and prevented.

A key element of fake news is emotional expression (Alonso et al., 2021). In most cases, various methods spread fake news to attract users' attention and mislead them to comment and forward. Fake news publishers generally utilize emotionally arousing tactics to drive users to respond with more exaggerated fabrications.

Emotional elements are consequently considered enrichment features for fake news detection. Previous studies by Wu et al (2020) found emotional correlations and semantic conflicts between news content and user comments. Furthermore, Zhang et al. (2021) found that user comments often

DOI: 10.4018/IJDWM.314585

\*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

included sentiment relating to the emotion of news content. Apart from focusing on the feeling of the news content, they explored the sentiment of the news comments and the difference between the generations.

Though crucial for detecting fake news, emotional information is still far from being fully used in these studies, calling for further explorations. First, when using emotional features of user comments, there is no screening, and often only the first few user comments are directly used (Zhang et al., 2021), which happens the same while using semantic features (Shu et al., 2019). In particular, for datasets such as Weibo (Ma et al., 2016), where the number of user comments is extremely large, no research has yet been conducted on selecting the most relevant user comments to detect fake news. Second, the correlation between user comment sentiment and news content sentiment is not fully considered (Zhang et al., 2021), and in existing models, the sentiment feature representations of news content and user comments are usually extracted separately as detector features. Finally, the sentiment features in user comments are not exploited to provide reasonable interpretability for fake news detection. While explainable fake news detection often starts from the semantic perspective (Shu et al., 2019) and the forwarding relationship (Lu et al., 2020), existing models that use emotional features for fake news, detection has not considered the emotional perspective to provide reasonable explanations.

To address the abovementioned issues, we propose an Emotion-Drive Interpretable fake news detection model (EDI) that selects user comments based on their emotional value, and utilize Convolutional Neural Networks (CNN) to extract sentiment representations of news content and user comments. Finally, the correlation between the emotional features of news content and user comments are learned through Co-attention, and the representation of user comments emotional features is learned through Attention. Finally, the weights of co-attention and attention provide interpretations

The principal contributions of this paper are: (1) We propose a novel interpretable fake news detection model, EDI, which selects user comments according to the sentiment value. In addition, it can fully learn the content of rumors and the sentiment representation of user comments for detection. (2) We present an interpretable structure in the EDI model, which adopts co-attention and attention mechanisms based on the emotional value of news posts and comments. (3) Experiments on Chinese and English fake news datasets show that our model also outperforms current state-of-the-art models, and the EDI also provides reasonable explanations, which are shown by a case study. The code of the EDI model can be accessed: <https://github.com/wj-gxy/EDI>

## **RELATED WORK**

### **Fake News Detection**

Recent research advances suggest dividing fake news detection into content-based, social context-based, and hybrid-based approaches (Shu, 2017; Zhou, 2019). Content-based approaches typically rely on textual, visual, and speech aspects. Textual features are usually semantic (Wu et al., 2018), emotional (Zhang et al., 2021), and writing styles (Potthast et al., 2017) that are common in fake news content. Visual features are taken from videos (Nataraj et al., 2019) or images (Wu et al., 2021) for fake news detection, as are speech features (Wu et al., 2021). Social context-based methods emphasize capturing features around social context, such as source-based (Tschitschek et al., 2018), post-based (Zhang, 2019; Wu, 2021), comment-based (Guo, 2018; Yang, 2022), user-based (Shu, 2019; Dou, 2021), and network (Shu, 2020; Bian, 2020). Hybrid-based approaches often fuse multimodal or multiple features for fake news detection (Wu, 2021; Lu, 2020). In recent years, research has also focused on explainable fake news detection (Shu, 2019; Lu, 2020; Yanget, 2019; Fu, 2022), aiming to improve fake news detection performance while highlighting evidence when news is predicted to be fake. In this paper, we capture emotional features from news content and user reviews and study interpretable fake news detection through emotion, combining the advantages of these two methods.

## Emotion Representation

Early text-based emotion representations relied on emotion dictionaries. For example, WordNet (Kamps et al., 2004), MPQA (Wiebe et al., 2005), and HowNet are all centralized and widely utilized emotion dictionaries. However, the feature extraction method based on the emotion dictionary cannot comprehensively obtain emotion representation, and it is more effective to use the deep learning method to obtain emotion embedding. Agrawal et al. (2018) learn emotion-rich word representations in product reviews but with a much smaller corpus. In the Emo2Vec model (Xu et al., 2018), emotion semantics are encoded into word-level representations of fixed-size real-valued vectors, and it is learned through six different emotion-related tasks under the multi-task learning framework. Seyeditabari et al. (2019) used emotion lexicons and mental models of basic emotions to assist in training. They found that fitting emotion models to pre-train word vectors could improve the performance of these models of emotion similarity metrics.

## Interpretable Fake News

With the wide application of deep learning algorithms in the fields of text classification (Yang et al., 2016), image classification (Nataraj et al., 2019), malware detection (Jha et al., 2020), recommendation system (Qi et al., 2022), and so on, although it plays an obvious role in various tasks, it is often criticized for the lack of explanation of results and opaque process. Users, developers, and regulators need the model to give reasonable explanations in tasks such as classification and recommendation, making the model transparent, credible, and consistent with ethical standards (Chakraborty et al., 2017). Therefore, it is significant for artificial intelligence models to provide reasonable explanations for the results.

Recently, interpretable deep learning methods have been widely used in network security (Xu et al., 2021), social networks (Lu et al., 2020), medical care (Pintelas et al., 2021), and other fields. The news released by social media platforms is often related to social news that is 'hot' and people's lives. Therefore, it is also necessary to describe the results after confirming that the news is true or false and providing the results to the government, social media platforms, and users. In interpretable fake news detection, recent research concentrates on the discovery of evidence to make the model interpretable or on the study of results using interpretable tools.

## PROBLEM STATEMENT

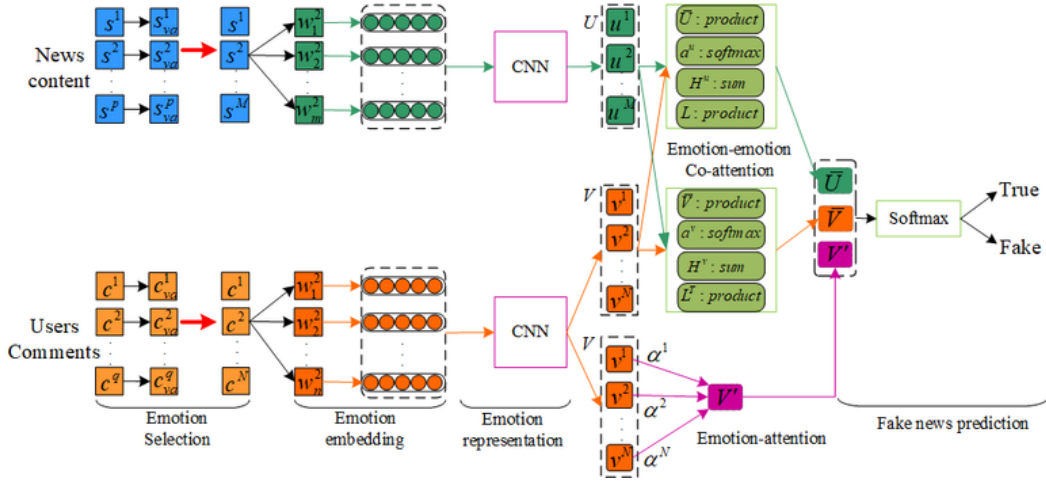
We describe the fake news detection task as follows. Let  $S = \{s^1, s^2, \dots, s^p\}$  be a news content, which contains  $p$  sentences, and each sentence  $s^i = \{w_1^i, w_2^i, \dots, w_m^i\}$  contains  $m$  words. There are a great number of utilizer comments, follow-up posts, and opinion contents following the published news. Let  $C = \{c^1, c^2, \dots, c^q\}$  be a set of  $q$  comments related to the news content  $S$ , where each user comment  $c^j = \{w_1^j, w_2^j, \dots, w_n^j\}$  contains  $n$  words. The fake news detection task is regarded as a binary classification task, and a binary label  $y \in \{0, 1\}$  is utilized to indicate the truthfulness of each piece of news. Our system needs to determine the label of the news: true or fake, and output the explanation of the decision-making process as the interpretation. The interpretability of sentences in news content shows how worthy of inspection they are, and the interpretability of user comments indicates the degree to which users believe the news is fake or true.

## MODEL

In this section, we introduce the proposed model, Emotion-Drive Interpretable Fake News Detection (EDI), which utilizes the emotional features of news and user comments to detect fake news. As shown in Figure 1, EDI consists of five components. First, emotion selection: obtaining the emotion value of

each sentence in the news text and user comments according to the emotion dictionary, and selecting emotion-rich ones. Second, emotion representation: obtaining emotion embedding vectors using the emotion feature pre-trained model, then feeding them into the Convolutional Neural Network (CNN) to catch the feature representation. Third, emotion attention: using the attention mechanism to learn attention weights that measure the importance of every comment and obtain attention content vectors of comments. Fourth, emotion-emotion co-attention: learning feature representations through the emotional correlation between news sentences and user comments. Fifth, fake news prediction: fake news prediction by concatenating the learned representations of the two sections.

Figure 1. The Model Architecture Proposed in this Paper Comprises Five Parts, Namely Emotion Selection, Emotion Representation, Emotion Attention, Emotion-Emotion Co-attention, and Fake News Prediction



## Emotion Selection

Emotional value, usually a positive or negative value, often reflects the degree of positive and negative polarity of text. The emotion dictionary can calculate it. We utilize an emotion dictionary to calculate the emotion value of each text for sentence  $s^i$  and user comment  $c^j$ . Consider not only the frequency of a word but also its degree of adverbs, affirmatives, and negatives in context. To begin with, compute the individual score of each word in the emotion dictionary, then match and compute the values of negative words and degree words through the existing sentiment dictionary. The following is the method for calculating the emotional value of each sentence:

$$s(w_t^i) = D(w_t^i) * neg(w_t^i, x) * deg(w_t^i, x) \quad (1)$$

$$D(w_t^i) = \begin{cases} 1, & \text{if } w_t^i \in D \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where  $D$  is the emotion dictionary,  $w_t^i$  is the vocabulary word in the sentence,  $x$  is the context range,  $neg(w_t^i, x)$  and  $deg(w_t^i, x)$  are the corresponding negative words values and degree adverb values:

$$neg(w_t^i, x) = \prod_{t-x}^{t-1} neg(w_t^i) \quad deg(w_t^i, x) = \prod_{t-x}^{t-1} deg(w_t^i) \quad (3)$$

It is easy to get the adverb of degree in the context of each word; then, we could obtain the emotion value of a sentence by adding up the calculated values of all words in that:

$$s_{va}^i = \sum_{t=1}^m s(w_t^i) \quad (4)$$

We select  $M$  sentences in the news and  $N$  comments for the next layer in our model according to their emotion value.

### Emotion Representation

We use the convolution neural network (CNN) (Kipf et al., 2017) to process the emotion embedding vector, then obtain the emotion representations. According to the latest process in emotion modelling, this paper adopts a method that incorporates emotion elements into the original model to obtain emotion embedding vectors. The method creates two sets of constraints based on the NRC emotion lexicon (Xu et al., 2008), one for words that have a positive relationship with an emotion (e.g., (kidnapping, sadness)) and another set for tracking each word that is opposite to that emotion (e.g., (abduction, joy)). By adding a new training stage, emotion vectors are obtained by fitting emotion information to pre-trained word vectors using emotion vocabulary and basic emotion vocabulary (Seyeditabari et al., 2019). We use the pre-trained Numberbatch (Speer et al., 2017) word vectors as the English word vector embedding method, since it considers the word vector similarity. The Weibo word vectors (Li et al., 2018) are used as our Chinese word embedding method. We show the statistical analysis of the positive and negative constraint dictionaries and the Chinese and English vector sizes in Table 1.

Table 1. The statistical analysis of each dictionary

Category	English	Chinese
Positive size	16502	22332
Negative size	16508	22012
Vocabulary size	516782	189600

CNN can be applied to image feature extraction (Kadry et al., 2022), text feature extraction (Alshubaily, 2021), and sentiment feature extraction. So, after we obtain the emotion embedding vector of every word, we feed them into the CNN model to learn the emotional features of news content and user comments separately. Finally, we adopt 1-D CNN to learn the emotion features of news content and user comments. Specifically, given a comment  $c^j$  with words  $w_t^j$ ,  $t \in \{1, 2, \dots, N\}$ , the emotion embedding of each word is  $e_t^j$ ,  $t \in \{1, 2, \dots, N\}$ , and the output emotion representation vector is obtained as follows:

$$\hat{v}^j = ReLU(W_f e_{t:t+\lambda-1}^j + b_f) \quad (5)$$

where  $W_f \in \mathbb{R}^{\lambda \times d}$  is the convolution filter and the matrix of learnable parameter, which is applied to a window of  $\lambda$  words to produce a new feature.  $b$  is the bias term, and  $ReLU()$  is the activation function. In order to catch emotional feature efficiently, we utilize two filters  $\lambda \in \{2, 3\}$ , respectively, and use Maxpooling to obtain the emotional features of each comment from them. Therefore, we could obtain all emotional features of every comment  $V = [v^1, v^2, \dots, v^N]$  and news content  $U = [u^1, u^2, \dots, u^M]$  based on above approach.

### Emotion Attention

Each user's comments could offer different importance for detecting fake news. Consequently, an attention mechanism (Vaswani et al., 2017) is used to measure the importance of each user comment.  $\alpha^j$  is the attention weight which indicates the importance of each comment  $v^j$ :

$$u^j = \tanh(W_w v^j + b_w) \quad (6)$$

$$\alpha^j = \frac{\exp(u^j u_w^T)}{\sum_{k=1}^N (u^k u_w^T)} \quad (7)$$

$$V' = \sum_{j=1}^N v^j \alpha^j \quad (8)$$

where  $u^j$  is obtained representation after feeding  $v^j$  to the full embedding layer,  $W_w$  is the trainable weight parameter,  $b_w$  is the bias term, and  $u_w$  is the weight parameter representing the word-level context vector. Then the emotional attention content vector of user comments can be computed as:

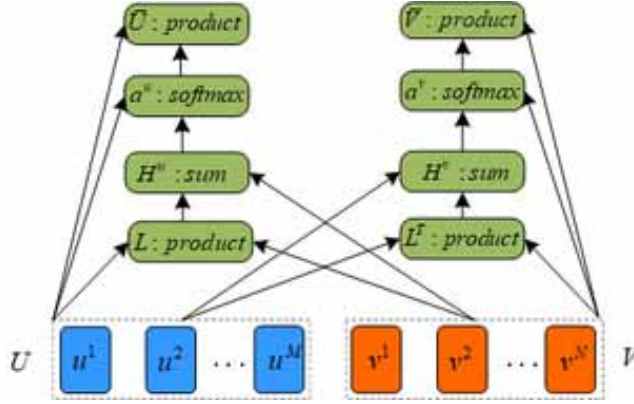
### Emotion-Emotion Co-Attention

We noticed that the difference in emotion between news content and user comments is an important feature of fake news detection (Zhang et al., 2021), so we are making relevant choices, especially those in which the emotion in news content is strongly correlated with the emotion of user comments. Not all sentences in news content are useful in emotion some sentences have high or low emotion value, but only to attract users' attention or comments. Therefore, news sentences play different roles in judging and explaining the truth of a piece of news. The same is true of user comments, which may contain important information explaining the truth of the news, but may also be less informative and noisy.

Specifically, we develop an emotion-emotion co-attention mechanism to capture the emotional affinity of news and comments and further learn the relevance of sentences and comments, as shown in Figure 2.

We first compute the affinity matrix  $F \in \mathbb{R}^{M \times N}$  utilizing the emotion representation of news sentences  $U = [u^1, u^2, \dots, u^M]$  and user comments  $V = [v^1, v^2, \dots, v^N]$ :

Figure 2. The Specific Process of the Co-Attention Model



$$F = \tanh(U^T W_{uv} V) \quad (9)$$

where  $W_{uv} \in \mathbb{R}^{2d \times 2d}$  is a learnable weight matrix. Next, we utilize the affinity matrix  $F$  to obtain the attention weights of sentences and comments, as follows:

$$H^u = \tanh(W_u U + (W_v V) F^T) \quad H^v = \tanh(W_v V + (W_u U) F) \quad (10)$$

where  $W_u$  and  $W_v$  are learnable weight parameters. The affinity matrix  $F$  transforms the emotional attention space of user comments into the emotional attention space of news words and transposes it into  $F^T$ . Then, the weight of emotional attention to news sentences and comments can be calculated by softmax function:

$$\alpha^u = \text{softmax}(w_{hu}^T H^u) \quad \alpha^v = \text{softmax}(w_{hv}^T H^v) \quad (11)$$

where  $w_{hu}$  and  $w_{hv}$  are learnable weight parameters. Then final emotional representation of news sentences  $\bar{U}$  and comments  $\bar{V}$  could be computed by the weighted sum of the emotional attention characteristics of sentences and user comments:

$$\bar{U} = \sum_{i=1}^M \alpha_i^u u^i \quad \bar{V} = \sum_{j=1}^N \alpha_j^v v^j \quad (12)$$

Where  $\alpha_i^u$ ,  $\alpha_j^v$  is the attention weight calculated from the above equation,  $u^i$ ,  $v^j$  are news sentences and user comments respectively,  $M$  and  $N$  are the number of news sentences and user comments selected by emotional value, respectively.

### Fake News Prediction

We select a certain number of news sentences and user comments through emotion value, learn the emotion features of news content and user comments through the CNN model, learn the weight of

user comments through the attention mechanism to obtain feature representation and use co-attention to learn sentences and comments based on emotional expression. Finally, we fuse the acquired emotional features to predict fake news. We aim to predict fake news using the emotional attention representation  $V'$  and emotional co-attention representation of news sentences  $\bar{U}$  and user comments  $\bar{V}$ . Therefore, we concatenate the acquired emotional features to predict fake news:

$$\hat{y} = \text{softma}\left(\left[\bar{U}, \bar{V}, V'\right]W_y + b_y\right) \quad (12)$$

where  $\hat{y}$  is the fake news prediction label,  $W_y$  is the weight parameter matrix, and  $b_y$  is the bias term.

## EXPERIMENT AND RESULT ANALYSIS

### Datasets

Emotion expression can be different in diverse languages and cultural backgrounds. Therefore, we experiment on Chinese (Weibo-16 and Weibo-20) and English datasets (Twitter), and their statistical analysis is shown in Table 2.

**Table 2. Statistics analysis of the three datasets. Max: indicates the maximum number of comments in all comments; Min: indicates the minimum number of comments in all comments; Ave: Average number of comments per news**

	Twitter	Weibo-16	Weibo-20
News	1154	3706	6362
Comments	13781	1874678	1983440
Fake	577	1355	3161
True	577	2351	3201
Max	42	44763	44763
Min	0	1	0
Ave	11.9	505.8	311.8

The Twitter dataset combined two classic datasets, Twitter15 and Twitter16 (Ma et al., 2017); we only use experimental data, which has labels as “true” and “fake,” and they include information about news content, retweeted comments, and labels.

The dataset Weibo-16 is a Chinese fake news detection benchmark dataset proposed by (Ma et al., 2016). Zhang (2021) considered the impact of repeated news on learning and evaluation, and they used a clustering algorithm based on text similarity to eliminate the repetition of fake news subsets.

The Weibo-20 dataset was proposed based on the Weibo-16 by Zhang et al. (2021). In this dataset there are 1355 Weibo-16 fake news items, and the news that Weibo Community Management Centre officially judged as fake information is further collected. In terms of genuine news items, the dataset has 2351 items of genuine news of Weibo-16 were kept, and 850 unique genuine news were collected.

### Baselines

We compare our proposed approach with several state-of-the-art methods, which are described:



- **RNN** (Ma et al., 2016): A rumor detection method based on GRU, which is also the first time to utilize a deep learning method for rumor detection.
- **CNN** (Kim, 2014): A method that uses convolutional neural networks to model news content and gets textual features of different granularities through multiple convolutional filters.
- **HAN** (Yang et al., 2016): A state-of-the-art model for fake news detection based on hierarchical attention neural networks that leverage word-level and sentence-level attention, respectively, to learn news content representations.
- **dEFEND** (Shu et al., 2019): A state-of-the-art interpretable fake news detection model that uses HAN to obtain semantic features of news content, uses a bidirectional gating unit network and attention mechanism to represent user comment features, and gets feature representations through co-attention for rumor fake news detection.
- **Dual emotion** (Zhang et al., 2021): A fake news detection model based on emotional features, which utilizes news content emotional features, user comment emotional features, and the emotional generation gap between the two to represent dual emotions, and connects news content semantic features for fake news detection.

In the Twitter dataset, the number of news sentences in the dEFEND model and the Dual emotion model is 1, the length is 32, 11 user comments are selected, and 32. The Dual emotion model uses Bi-GRU to extract text features. The other models refer to the original parameters.

In the Weibo-16 and Weibo-20 datasets, the number of news sentences in the dEFEND and Dual emotion models is 1, the length is 64, and we selected 100 user comments with a length of 32. The Dual emotion model uses Bi-GRU to extract text features. The other models refer to the original parameters.

In the EDI model, we choose different sentence lengths, the number of user comments, and learning rates depending on the dataset. We show the settings in Table 3.

**Table 3. The details of the parameters of EDI**

	Twitter	Weibo-16	Weibo-20
Sentence count	1	1	1
Sentence length	32	64	64
Comment count	12	300	300
Comment selection	10	100	100
Comment length	32	32	32
Embedding dimension	300	300	300
Batch size	32	32	32
Learning rate	0.001	0.005	0.005
L2 regularization	0.001	0.001	0.001

## Model Comparison Results

The evaluation metrics include Accuracy, Precision, Recall, and F1. We divided the experimental data into a training set, a validation set, and a test set according to the ratio of 6:2:2. We repeat the experiment five times to take the average test set results as the final result. The experimental results are shown in Tables 4 and 5, respectively.

Tables 4 and 5 show that the proposed EDI model outperforms the baseline model in three datasets. The accuracy and F1 score on Twitter are improved by 1.1% and 1.8%, respectively. In Weibo-16,

Table 4. Main results from the Twitter data. The best model and competitor are highlighted with bold and underlining, respectively

Method	Twitter			
	Acc	Pre	Rec	F1
RNN	0.752	0.753	0.747	0.747
CNN	0.765	0.763	0.752	0.752
HAN	0.787	0.790	0.788	0.786
dEFEND	0.828	0.830	0.829	0.828
Dual emotion	<u>0.838</u>	<u>0.835</u>	<u>0.833</u>	<u>0.831</u>
EDI	<b>0.849</b>	<b>0.854</b>	<b>0.851</b>	<b>0.849</b>

Table 5. Main results in Weibo-16 and Weibo-20. The best model and competitor are highlighted with bold and underlining, respectively

Method	Weibo-16			Weibo-20				
	Acc	Pre	F1	Acc	Pre	Rec	F1	
RNN	0.701	0.700	0.653	0.652	0.713	0.713	0.713	0.713
CNN	0.716	0.706	0.665	0.671	0.775	0.775	0.775	0.775
HAN	0.720	0.717	0.664	0.669	0.721	0.721	0.721	0.721
dEFEND	<u>0.821</u>	0.817	<u>0.822</u>	<u>0.821</u>	0.843	0.843	0.840	0.843
Dual emotion	0.819	<u>0.826</u>	0.784	0.796	<u>0.863</u>	<u>0.865</u>	<u>0.862</u>	<u>0.862</u>
EDI	<b>0.888</b>	<b>0.886</b>	<b>0.873</b>	<b>0.879</b>	<b>0.910</b>	<b>0.908</b>	<b>0.907</b>	<b>0.907</b>

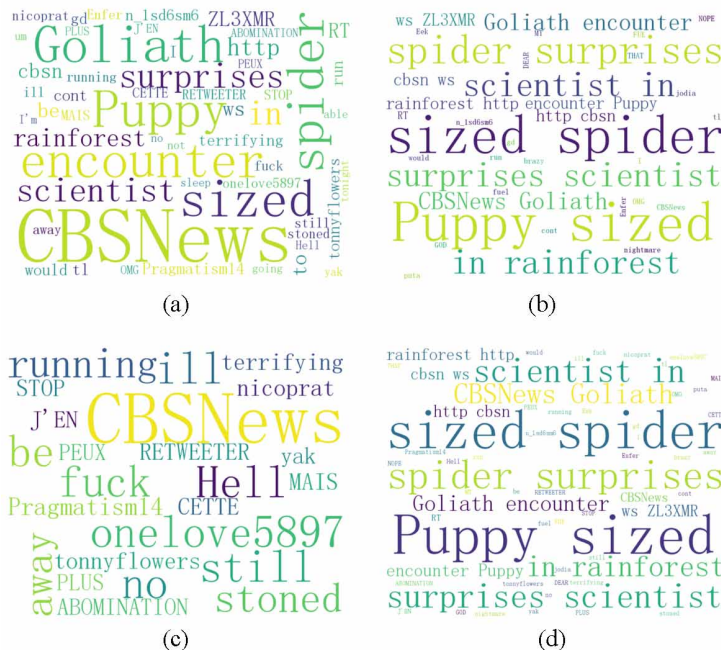
the accuracy is improved by 6.9%, and the F1 score is improved by 8.3%. In Weibo-20, accuracy is improved by 4.7%, and the F1 score is improved by 4.5%. Comparing the improvement results of the three datasets shows that dataset Weibo-16 has the most rapid improvement in accuracy and F1 score. The analysis also shows that dataset Weibo-16 has the largest average user comment and stronger selectivity, and the effect of selecting comments is the best. On the contrary, the Twitter dataset has the least accuracy and F1 score improvement. The analysis shows that the number of comments made by comment users is 11.9, which means there is little choice in the emotional selection stage.

In addition, we represent other result analyses. First, using news content as an input feature alone is not as effective as the method based on mixed features. For example, RNN, CNN, and HAN are not as effective as Defend, dual Emotion, and our model in terms of accuracy. Second, the Dual Emotion model, which was then the dEFEND model, uses emotional features as enhancement. The dEFEND model only uses semantic features of news content and user comments, while the Dual Emotion model uses emotional features of news content and user comments as enhanced features based on semantic features of news content, which can fully advance features and obtain better results. Third, although EDI and dEFEND are both models based on co-attention, in the EDI model, the module that selects comments based on emotion value is added to filter user comments in advance. The module of attention mechanism is used to select comments independently to get better effects. Finally, EDI and Dual Emotion are fake news detection models based on emotional features. However, Dual Emotion only uses emotional features as enhanced features without considering comment selection and takes the first 100 comments as input, resulting in insufficient emotional features. On the other hand, emotion representation in EDI is added, and the learned emotional traits can significantly improve performance. The ablation study section elaborates on the influence of single modules on EDI model performance.

### User reviews Selection Comparison

From Figure 3, we can see the difference between the four sub-figures. In Figure 3a, there are comments without emotion value, and the information is the content of the first ten comments. In Figure 3b, for the ten comments with the maximum emotion value, we can see that some words, such as “surprise” and “puppy” contain positive words, but there is no word like “CBSNEWS.” For Figure 3c, we have ten comments with the minimum emotion value. Some offensive words like “fuck” and “no” can be seen, but words like “surprise” are lost. For Figure 3d, the five comments with the maximum emotion value and the five with the minimum emotion value, we can clearly see that all the above words are included, and this key information can be well included.

Figure 3. Select the Effect of Comments on Twitter Based on Emotional Value. (a): Top 10 Comments Without Selection, (b): 10 Comments with Maximum Emotional Value, (c): 10 Comments with Minimum Emotional Value, (d): 5 Comments with Maximum Emotional Value and Minimum Emotional Value

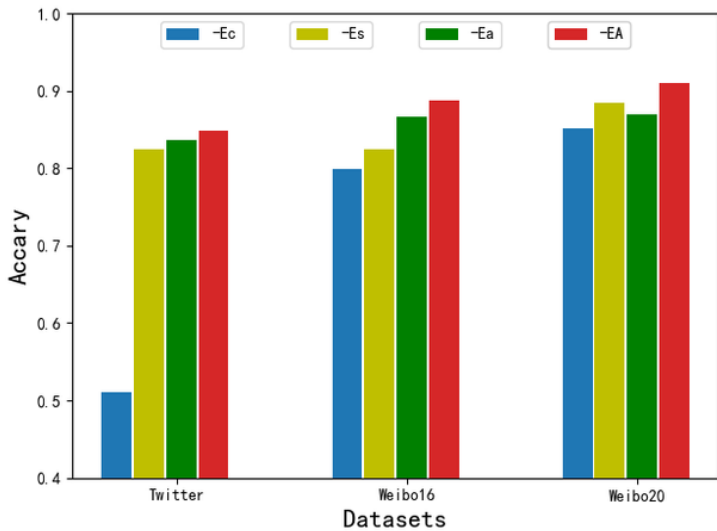


## Component Analysis

We conducted experiments on the removal of each module of the model to reflect the contribution of each module to the model. The sub-model “-Es” represents the model after removing the emotional selection, the sub-model “-Ec” represents the model with the emotion-emotion co-attention removed, the sub-model “-Ea” represents the model after removing the emotional attention, and “EA” represents the EDI model. The results are shown in Figure 4.

In Figure 4, we can see that the model’s accuracy decreases when any model component is removed. However, the largest drop in accuracy can be seen when the emotion-emotion co-attention component (i.e., the ‘Ec’ sub-model) is removed, and the performance is most pronounced on the Twitter dataset, indicating that relying on user comments alone for fake news detection is weak. In addition, when we remove the emotion selection component (i.e., the ‘Es’ sub-model), the model’s accuracy decreases significantly and is most pronounced on the Weibo-16 dataset. This observation occurs as the Weibo-16 dataset has the largest average number of user comments in the dataset and requires the selection of comments for the next feature extraction. Finally, when the emotion attention component is removed, the accuracy drops even more than when the emotion selection module is removed on the Weibo-20 dataset.

Figure 4. EDI Ablation Analysis in Accuracy



### The Case Study of Interpretability

By showing the location of the distribution of attention weights, evidence for predicting fake news can be revealed. The co-attention weights and attention mechanism weights from the above enable our model to provide reasonable explanations from the perspective of user reviews. The following is a case analysis of the co-attention weight and the attention weight in the model. Before that, we first select reviews by emotion value, and the results are shown in Figure 5, and the news content is “BREAKING: Malaysia Airlines passenger ‘shot down’ near Russian border in Ukraine URL URL\n”.

#### Interpretability of Co-Attention

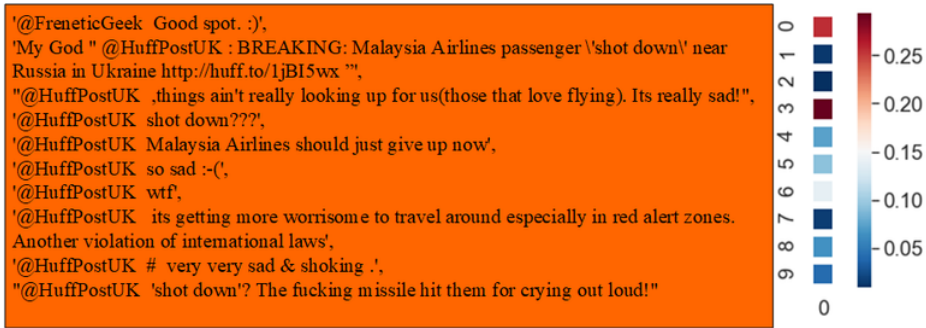
We use co-attention weights to reveal news content and attention sentences in user reviews. The number of fake news sentences is set to 1, so the focus on fake news is always the only sentence of the news content. User comments are selected by weight distribution. The weight distribution of final user comments is shown in Figure 6, and the weights correspond to user comments one by one.

In Figure 6, the first sentence “@FreneticGeek Good spot :)” and the fourth sentence, “@HuffPostUK shot down???” in the user comments, have the most weight. In addition, these two user comments are closely related to the news content, for example, using the word “shot down,” as in the news content.

Figure 5. Emotional Value Selects the Result of the Comment



Figure 6. The Explanation for Comments is Based on the Co-Attention Mechanism. heat maps show the Co-Attention Weights, and a Higher Attention Weight Represents how the Comment is Related to the Post and is Important to the Decision-Making Process

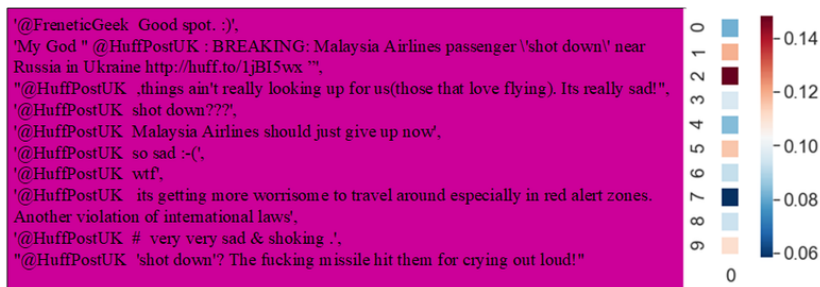


### Interpretability of Attention

When using attention to select user comments, we do not consider the news content and only assign the weights of user comments, so the corresponding distribution of attention weights in user comments is shown in Figure 7, where user comments and heatmaps correspond one-to-one.

In Figure 7, the second “My God \” @HuffPostUK: BREAKING: Malaysia Airlines passenger ‘shot down’ near Russia in Ukraine <https://huff.to/1jBI5wx>” and third “@HuffPostUK, things ain’t

Figure 7. The Explanation for Comments is Based on the Attention Mechanism. Heat Maps Show the Attention Weights, and a Higher Attention Weight Represents that the Comment is Important to the Decision-Making Process



really looking up for us(those that love flying). Its really sad!” sentences are given the most weight in the user comments. The words “BREAKING” and “sad” in these two sentences are used to determine whether the news content is true or false.

Comparing the distribution of the two weights on user comments, we can find that the choices of the two kinds of attention have a sizeable gap and similar choices. The first two choices for synergistic attention weights are “@FreneticGeek Good spot.:)” and “@HuffPostUK shot down???” and the first two comments for attention weights are “My God \” @HuffPostUK: BREAKING: Malaysia Airlines passenger ‘shot down’ near Russia in Ukraine <https://huff.to/1jBI5wx>” and “@HuffPostUK, things ain’t really looking up for us(those that love flying). Its really sad!” We can find that what synergistic attention chooses is closely related to the news content, as can be seen from the “spot” and “shot down”, the attention chooses some sentences of shock and sadness, we can see from the word “God”, “BREAKING” and “sad” are seen. We found that the two attentions did not make any choice between “@HuffPostUK Malaysia Airlines should just give up now” and “@HuffPostUK its getting more worrisome to travel around especially in red alert zones. Another violation of international laws”.

## CONCLUSION

In this paper, we propose a model for fake news detection through emotional features, which explores further the following three aspects: (1) user comment selection, (2) emotional correlation between news content and user comments, and (3) interpretability through emotion. First, we select user comments as input through emotion value and use CNN to get emotion features. To provide more explanations, co-attention is applied to obtain relevant representations of news content and user comments, and attention weights are developed to obtain weight representations of user comments and fuse features for fake news detection. The experimental results show the model has high detection accuracy and reasonable interpretability. However, some sentences contain words with opposite emotions, which offset the overall emotion of the sentence. In future work, we will combine emotion and semantics with studying fake news detection, which will be an interesting topic.

## COMPETING INTERESTS

The authors have declared that there is no conflict of interest regarding the publication of this paper.

## REFERENCES

- Agrawal, A., An, A., & Papagelis, M. (2018). Learning emotion-enriched word representations. *Proceedings of the 27th International Conference on Computational Linguistics*, 950–961. <https://aclanthology.org/C18-1081>
- Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *The Journal of Economic Perspectives*, 31(2), 211–236. doi:10.1257/jep.31.2.211
- Alonso, M. A., Vilares, D., Gómez-Rodríguez, C., & Vilares, J. (2021). Sentiment analysis for fake news detection. *Electronics (Basel)*, 10(11), 1348. doi:10.3390/electronics10111348
- Alshubaily, I. (2021). *TextCNN with attention for text classification*. .10.48550/arXiv.2108.01921
- Bian, T., Xiao, X., Xu, T., Zhao, P., Huang, W., Rong, Y., & Huang, J. (2020). Rumor detection on social media with bi-directional graph convolutional networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(1), 549–556. doi:10.1609/aaai.v34i01.5393
- Chakraborty, S., Tomsett, R., Raghavendra, R., Harborne, D., Alzantot, M., Cerutti, F., ... Gurram, P. (2017). Interpretability of deep learning models: A survey of results. *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, 1–6. 10.1109/UIC-ATC.2017.8397411
- Dou, Y., Shu, K., Xia, C., Yu, P. S., & Sun, L. (2021). *User preference-aware fake news detection*. 10.48550/arXiv.2104.12259
- Fu, D., Ban, Y., Tong, H., Maciejewski, R., & He, J. (2022). *DISCO: Comprehensive and explainable disinformation detection*. 10.48550/arXiv.2203.04928
- Guo, H., Cao, J., Zhang, Y., Guo, J., & Li, J. (2018). Rumor detection with hierarchical social attention network. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 943–951. doi:10.1145/3269206.3271709
- Haq, E., Tyson, G., Lee, L.-H., Braud, T., & Hui, P. (2022). *Twitter dataset for 2022 Russo-Ukrainian Crisis*. doi:10.48550/arXiv.2203.02955
- Jha, S., Prashar, D., Long, H. V., & Taniar, D. (2020). Recurrent neural network for detecting malware. *Computers & Security*, 99, 102037. doi:10.1016/j.cose.2020.102037
- Kadry, S., Rajinikanth, V., Taniar, D., Damasevicius, R., & Blanco-Valencia, X. P. (2022). Automated segmentation of leukocyte from hematological images—A study using various CNN schemes. *The Journal of Supercomputing*, 78(5), 6974–6994. doi:10.1007/s11227-021-04125-4
- Kamps, J., Marx, M., Mokken, R. J., De Rijke, M., & Associates. (2004). Using WordNet to measure semantic orientations of adjectives. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. <http://www.lrec-conf.org/proceedings/lrec2004/pdf/734.pdf>
- Kim, Y. (2014). Convolutional neural networks for sentence classification. *EMNLP*.
- Kipf, T. N., & Welling, M. (2017). *Semi-supervised classification with graph convolutional networks*. doi:10.48550/arXiv.1609.02907
- Li, S., Zhao, Z., Hu, R., Li, W., Liu, T., & Du, X. (2018). Analogical reasoning on Chinese morphological and semantic relations. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Volume 2: Short Papers* (pp. 138–143). Association for Computational Linguistics. doi:10.18653/v1/P18-2023
- Lu, Y.-J., & Li, C.-T. (2020). GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 505–514. doi:10.18653/v1/2020.acl-main.48
- Ma, J., Gao, W., Mitra, P., Kwon, S., & Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks. *International Joint Conference on Artificial Intelligence*, 3818–3824.

- Ma, J., Gao, W., & Wong, K.-F. (2017). Detect rumors in microblog posts using propagation structure via kernel learning. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Volume 1: Long Papers* (pp. 708–717). Association for Computational Linguistics. doi:10.18653/v1/P17-1066
- Nataraj, L., Mohammed, T. M., Chandrasekaran, S., Flenner, A., Bappy, J. H., Roy-Chowdhury, A. K., & Manjunath, B. S. (2019). *Detecting GAN generated fake images using co-occurrence matrices*. 10.48550/arXiv.1903.06836
- Pintelas, E. G., Liaskos, M., Livieris, I. E., Kotsiantis, S., & Pintelas, P. E. (2021). A novel explainable image classification framework: Case study on Skin cancer and Plant disease prediction. *Neural Computing & Applications*, 33(22), 15171–15189. doi:10.1007/s00521-021-06141-0
- Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2017). *A stylometric inquiry into hyperpartisan and fake news*. 10.48550/arXiv.1702.05638
- Qi, T., Wu, F., Wu, C., & Huang, Y. (2022). *FUM: Fine-grained and fast user modeling for news recommendation*. 10.48550/arXiv.2204.04727
- Seyeditabari, A., Tabari, N., Gholizade, S., & Zadrozny, W. (2019). *Emotional embeddings: refining word embeddings to capture emotional content of words*. 10.48550/arXiv.1906.00112
- Shu, K., Cui, L., Wang, S., Lee, D., & Liu, H. (2019). dEFEND: Explainable Fake News Detection. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 395–405. doi:10.1145/3292500.3330935
- Shu, K., Mahudeswaran, D., Wang, S., & Liu, H. (2020). Hierarchical propagation networks for fake news detection: Investigation and exploitation. *Proceedings of the 14th International AAAI Conference on Web and Social Media, ICWSM 2020*, 626–637. doi:10.1609/icwsml.v14i1.7329
- Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *SIGKDD Explorations*, 19(1), 22–36. doi:10.1145/3137597.3137600
- Shu, K., Zhou, X., Wang, S., Zafarani, R., & Liu, H. (2019). *The role of user profile for fake news detection*. 10.48550/arXiv.1904.13355
- Speer, R., Chin, J., & Havasi, C. (2017). ConceptNet 5.5: An open multilingual graph of general knowledge. *AAAI Conference on Artificial Intelligence Thirty-First AAAI Conference on Artificial Intelligence*, 4444–4451. doi:10.1609/aaai.v31i1.11164
- The Lancet Infectious Diseases. (2020). The COVID-19 infodemic. *The Lancet. Infectious Diseases*, 20(8), 875. doi:10.1016/S1473-3099(20)30565-X PMID:32687807
- Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A., & Krause, A. (2018). Fake news detection in social networks via crowd signals. *Companion Proceedings of the Web Conference 2018*, 517–524.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All You Need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010.
- Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2), 165–210. doi:10.1007/s10579-005-7880-9
- Wu, H., Kuo, H.-C., Zheng, N., Hung, K.-H., Lee, H.-Y., Tsao, Y., Wang, H.-M., & Meng, H. (2022). Partially fake audio detection by self-attention-based fake span discovery. *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 9236–9240.
- Wu, L., Rao, Y., Lan, Y., Sun, L., & Qi, Z. (2021). Unified dual-view cognitive model for interpretable claim verification. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 59–68. doi:10.18653/v1/2021.acl-long.5
- Wu, L., Rao, Y., Yu, H., Wang, Y., & Nazir, A. (2018). False information detection on social media via a hybrid deep model. *International Conference on Social Informatics*, 323–333.
- Wu, L., & Yuan, R. (2020). Adaptive interaction fusion networks for fake news detection. *ECAI*, 2220–2227.



- Wu, Y., Zhan, P., Zhang, Y., Wang, L., & Xu, Z. (2021). *Multimodal fusion with co-attention networks for fake news detection*. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics. doi:10.18653/v1/2021.findings-acl.226
- Xu, L., Lin, H., Pan, Y., Ren, H., & Chen, J. (2008). Constructing the affective lexicon ontology. *Journal of the China Society for Scientific and Technical Information*, 27, 180–185.
- Xu, P., Madotto, A., Wu, C.-S., Park, J. H., & Fung, P. (2018). *Emo2vec: Learning generalized emotion representation by multi-task training*. 10.48550/arXiv.1809.04505
- Xu, X., Zheng, Q., Yan, Z., Fan, M., Jia, A., & Liu, T. (2021). Interpretation-enabled software reuse detection based on a multi-level birthmark model. *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, 873–884.
- Yang, F., Pentyala, S. K., Mohseni, S., Du, M., Yuan, H., Linder, R., & Hu, X. et al. (2019). XFake: Explainable Fake News Detector with Visualizations. *The World Wide Web Conference on - WWW '19*, 3600–3604. doi:10.1145/3308558.3314119
- Yang, Y., Wang, Y., Wang, L., & Meng, J. (2022). PostCom2DR: Utilizing information from post and comments to detect rumors. *Expert Systems with Applications*, 189(c), 116071. doi:10.1016/j.eswa.2021.116071
- Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480–1489. doi:10.18653/v1/N16-1174
- Zhang, Q., Lipani, A., Liang, S., & Yilmaz, E. (2019). Reply-aided detection of misinformation via bayesian deep learning. *The World Wide Web Conference*, 2333–2343. doi:10.1145/3308558.3313718
- Zhang, X., Cao, J., Li, X., Sheng, Q., Zhong, L., & Shu, K. (2021). Mining dual emotion for fake news detection. *Proceedings of the Web Conference 2021*, 3465–3476. doi:10.1145/3442381.3450004
- Zhou, X., Zafarani, R., Shu, K., & Liu, H. (2019). Fake news: Fundamental theories, detection strategies and challenges. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 836–837.

Xiaoyi Ge is missing a biography. Ming Shu Zhang was born in Henan, China, in 1978. He received a Ph.D. degree in cryptography from Xi'an Jiaotong University, Xi'an, China, in 2016. He has been with the Key Laboratory of Network and Information Security under PAP as an associate professor. His research lies in cybersecurity, data mining, and social computing with applications such as fake news and rumor.

Xu An Wang is currently a professor in Engineering University of CAPF. His main interests are information and cloud security, cryptography, etc. He has published more than 100 papers in these field. He also server the TPC co-chair for INCOS 2016/3PGCIC-2017 and workshop co-chair or TPC member for several other conferences. He is editor member or review board member of some journals like IJGUC/IJITWE/IJTHI/IJCCE.

Jia Liu graduated from Shanghai Jiao Tong University, majoring in machine learning and deep learning.