IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 15, 2020, accepted February 2, 2021, date of publication February 11, 2021, date of current version February 24, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3058809

Unsupervised Fake News Detection Based on Autoencoder

DUN LI, HAIMEI GUO, ZHENFEI WANG, AND ZHIYUN ZHENG
College of Information Engineering, Zhengzhou University, Zhengzhou 450001, China

Corresponding author: Zhiyun Zheng (iezyzheng@zzu.edu.cn)

This work was supported in part by the National Social Science Foundation Project under Grant 17BXW065.

ABSTRACT With the development of social networks, the spread of fake news brings great negative effects to people's daily life, and even causes social panic. Fake news can be regarded as an anomaly on social networks, and autoencoder can be used as the basic unsupervised learning method. So, an unsupervised fake news detection method based on autoencoder (UFNDA) is proposed. This paper firstly considers some forms of news in social networks, integrates the text content, images, propagation, and user information of publishing news to improve the performance of fake news detection. Next, to obtain the hidden information and internal relationship between features, Bidirectional GRU(Bi-GRU) layer and Self-Attention layer are added into the autoencoder, and then reconstruct residual to detect fake news. The experimental results compared with the existence of other four methods, on two real-world datasets, show that UFNDA obtains the more positive results.

INDEX TERMS Social networks, fake news detection, unsupervised learning, self-attention, autoencoder.

I. INTRODUCTION

With the widespread use of mobile phones, social networks users can easily share information with others, keep in touch, and learn hot news trend anytime, anywhere. However, some news are misleading, and their reliability is doubtful. Fake news defined by the authors in [1] is ''news articles that are intentionally and verifiably false, and could mislead readers.'' the authors in [2] proposed another definition, ''False information spread under the guise of being authentic news usually spread through news outlets or internet with an intention to gain politically or financially, increase readership, biased public opinion.'' It is the continuous content and dense distribution of fake news cause the major negative impacts on society. For example, by February 8th 2021, there are confirmed over 106 million COVID-19 cases and have been over 2 million COVID-19 related deaths in the world.[1] However, ''COVID-19 is a Hoax'' still spreads on social networks.[2] Another example, whenever a major hurricane start bearing down upon the America, fake news and pictures emerge in large quantities, and it is hard to distinguish real or fake from them.[3] There have been reported that fake news has cost the stock market 39 billion USD annually.[4] Manual identification of fake news requires experts in various fields and a huge workload. Therefore, a tool is needed to detect fake news. But there are vast volumes of news, and creating such a tool is very challenging.

As the technology of artificial intelligence and machine learning advances, supervised learning gradually apply to fake news detection now and win very great effectiveness. Four aspects are mainly utilized by detecting fake news on social networks: the text content of news, social context information, propagation information, and multimedia information. Text content and context information of news was used to detect fake news [3]–[5]. In [6] and [7], the authors evaluated the credibility of target articles by network structure. 23 kinds of supervised machine learning classification methods were evaluated on existing public datasets [8]. References [9], [10] adopted the method of Convolutional Neural Networks (CNN) to detect fake news, which showed quite successful and positive results. Although such methods achieved great results, the methods need large amounts of

The associate editor coordinating the review of this manuscript and approving it for publication was Santhosh Kumar Gopalan.

[1]https://en.wikipedia.org/wiki/Template:COVID-19_pandemic_data

[2]https://www.snopes.com/ap/2020/12/04/nevada-doctors-selfie-used-to-claim-covid-19-is-a-hoax/

[3]https://www.king5.com/article/news/verify/verify-fake-hurricane-photos/77-3a8ee689-e47e-46ac-a99d-4f5fb9295bc3

[4]https://www.institutionalinvestor.com/article/b1j2ttw22xf7n6/Fake-News-Creates-Real-Losses

labeled data to train the model. However, such a way is time-consuming and a significant amount of work. Therefore, some researchers apply unsupervised methods to fake news detection.

The authors in [11] proposed one method based on text content for unsupervised fake news detection, which used tensor to classify news according to hidden content of news, but it ignored user information for publishing news. Regarding the actual state of news and user reputation as latent random variables, and proposed a fold Gibbs method to infer the truth of news, but this method ignored multimedia information of news [12]. The authors in [13] utilized the technology of semantic similarity and transfer learning to detect the truth of news, but this way neglected the social context information of news. In [14], the authors proposed an approach based on graphic using users' behavior to detect unsupervised fake news, but this approach lose sight of text content of news. Some studies considered the early detection of fake news but neglected the detection after news propagate[15].

Currently, the technology of autoencoder mainly apply to many fields such as industrial system anomaly detection [16]–[18], medical diagnose lesions [19]–[21], image processing [22]–[24], etc. Because of its features of reproducing input data as far as possible, autoencoder is used to anomaly detection, which obtains a good result. So, this paper combines autoencoder and unsupervised learning to propose a new method based on autoencoder to detect fake news. The main contributions of our paper are summarized as follows:

1) To make full use of news in social networks, this paper fuses four kinds of features, including text content, images, propagation information, and user information of publishing news.

2) Regard fake news as an anomaly on social networks and make use of autoencoder as the basic unsupervised learning method.

3) Propose an improved method based on autoencoder for unsupervised fake news detection, and design comparative experiments to verify the effectiveness of the proposed method.

4) Evaluate the proposed method performance of early fake news detection.

5) Evaluate the effectiveness of splicing fusion.

The reminder of the paper is organized as follows. In Section II, we give a brief review of related works about fake news detection. Afterward, we introduce the architecture of our unsupervised fake news detection method based on autoencoder, namely UFNDA, and provide the details of how we construct the model based on autoencoder and the proposed method UFNDA algorithm in Section III. Then, Section IV evaluates our method with dedicated experiments and this paper ends with the conclusion in Section V.

II. RELATED WORK

At present, the methods of fake news detection mainly based on feature rules, machine learning, and deep learning. For the methods based upon feature rules, In [25], the authors collected digital information about the Chilean earthquake on Twitter and obtained that it is practicable to detect fake information through tweets. The authors in [26] analyzed the time, structure and language features of tweets and obtained the fake news detection accuracy of 92%. The authors in [27] analyzed the blog posts, images, and visual features in Sina Weibo and obtained the fake news detection accuracy of 83.6%. In [28], the authors analyzed news based on features such as content, users, and emotions, and obtained the fake news detection accuracy of 91% and the early fake news detection accuracy of 77%.

For the methods based upon machine learning, In [29], the authors used TF-IDF as the feature extraction method and linear SVM as the classifier, which improved the accuracy of fake news detection to 92%. The authors in [30] proposed a method based on text content and structure and used RF to analyze, which concluded that the content-based model showed bias detection. The authors in [31] proposed a method based on publisher, news content, and user features to train seven classifiers, and the accuracy of fake news detection was as high as 91.6%.

As for the methods based upon deep learning, In [32] authors proposed a text and image based on CNN model, and the accuracy of fake news detection was 92.2%. In [33], the authors proposed a Bi-LSTM model based on text summarization features, and obtained a fake news detection accuracy of 93.1%. In [34] authors proposed a Deep Convolutional Neural Network model, which obtained a fake news detection accuracy of 98.36%. The authors in [35] proposed a model detecting fake information automatically, and obtained the accuracy of 74%. In [36], the authors proposed a fake news automatic detection model, and showed that pre-trained deep learning models such as BERT, XLNet, and RoBERTa performed better than machine learning models such as SVM, RF, and XGBoost, and the accuracy of fake news detection was up to 98%.

Most of the above studies discern fake news by supervised learning. Considering that a large amount of news lacks manual labels, we integrate the text content features, propagation features, image features, and features of users who publish the news in social networks, and propose a method based on autoencoder to achieve the detection of unsupervised fake news in social networks.

III. METHOD

In social networks, most of software platforms provide many easy-to-use functionalities for internet users, which makes it more convenient to interact with other netizens by various information. There are many factors for netizens to infer the credibility of news, such as text content, comments, likes, embedded images and videos, authoritarian standpoints.

Autoencoder widely apply to anomaly detection. This paper treats fake news as an abnormal data in social networks and proposes a method based on autoencoder for unsupervised fake news detection with various features, namely UFNDA. After training UFNDA with real news data,

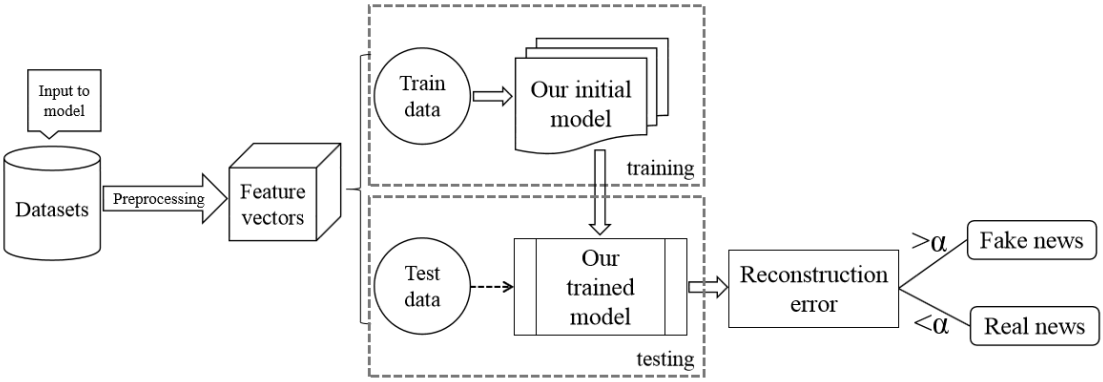


FIGURE 1. The flow of the proposed method.

the UFNDA has the ability to detect fake news. By testing the test set, to obtain the reconstructed residual to distinguish real news and fake news.

UFNDA mainly includes feature extraction, feature fusion, and the model based on autoencoder. Figure 1 is the flow of the proposed method.

A. FEATURES EXTRACTION AND FUSION

In this paper, the features extracted from social networks news are text content features, propagation features, image features, and user features.

1) TEXT CONTENT FEATURES

For making news appear more authentic, in social networks, users who publishing news often use external links to help explain news. To make the spread more widely, news is often participated in multiple topics and mentioned multiple users. News is often used more modal images, personal pronouns, capitalization, emotional words, etc. to make it more attractive. Therefore, this paper detects the authenticity of news in social networks by extracting features such as keywords, symbols, and links.

Here, documents set D describing a series of news and containing N documents is indicated as $D = \{d_1, d_2, \ldots, d_N\}$. For example, the property of positive emotion word quantity F_{C1} can be gained by function f_{C1}, as follows, where document d_1 contains l_1 words, positive sentiment dictionary is denoted as PSD, $word_j$ is the jth word in d_1.

$$f_{C1} = lambda\ word_j, PSD: \quad 1\ if\ word_j\ in\ PSD\ else\ 0. \tag{1}$$

$$f_{C1}(d_1) = \sum_{0}^{l_1} f_{C1}(word_j) \tag{2}$$

$$F_{C1} = [f_{C1}(d_1), f_{C1}(d_2), \ldots, f_{C1}(d_N)]^T \tag{3}$$

Another example, the property of negative emotion word quantity F_{C2} can be gained by function f_{C2}, as follows, where document d_2 contains l_2 words, negative sentiment dictionary is denoted as NSD.

$$f_{C2} = lambda\ word_j, NSD: \quad 1\ if\ word_j\ in\ NSD\ else\ 0. \tag{4}$$

$$f_{C2}(d_1) = \sum_{0}^{l_1} f_{C2}(word_j) \tag{5}$$

$$F_{C2} = [f_{C2}(d_1), f_{C2}(d_2), \ldots, f_{C2}(d_N)]^T \tag{6}$$

Likewise, other detail properties can be obtained by other functions, and the text content features can be indicated as matrix A_1, as the following:

$$A_1 = \begin{bmatrix} f_{C1}(d_1) & f_{C2}(d_1) & \ldots & f_{Cm}(d_1) \\ f_{C1}(d_2) & f_{C2}(d_2) & \ldots & f_{Cm}(d_2) \\ \ldots & \ldots & \ldots & \ldots \\ f_{C1}(d_N) & f_{C2}(d_N) & \ldots & f_{Cm}(d_N) \end{bmatrix}$$

where m is the dimension of these features, f_{Ci} indicates the function of extracting property, $i = 1, \ldots, m$.

2) PROPAGATION FEATURES

If content of news differs greatly from common sense, is more controversial, and is more sensational than ordinary news, it is more likely to cause users to forward it. Therefore, we detect the credibility of propagating content by extracting the number of retweet, comments, and likes. The propagation features can be indicated as matrix A_2, as the following:

$$A_2 = \begin{bmatrix} f_{P1}(d_1) & f_{P2}(d_1) & \ldots & f_{Pn}(d_1) \\ f_{P1}(d_2) & f_{P2}(d_2) & \ldots & f_{Pn}(d_2) \\ \ldots & \ldots & \ldots & \ldots \\ f_{P1}(d_N) & f_{P2}(d_N) & \ldots & f_{Pn}(d_N) \end{bmatrix}$$

where n is the dimension of these features, f_{Pj} indicates the function of extracting property, $j = 1, \ldots, n$.

3) IMAGE FEATURES

Modified images are often combined with news to make them appear more authentic. Therefore, we detect the reliability of news by extracting whether the images quoted in news are tampered. The image features can be indicated as matrix A_3, as the following:

$$A_3 = \begin{bmatrix} f_{I1}(d_1) & f_{I2}(d_1) & \dots & f_{Ip}(d_1) \\ f_{I1}(d_2) & f_{I2}(d_2) & \dots & f_{Ip}(d_2) \\ \dots & \dots & \dots & \dots \\ f_{I1}(d_N) & f_{I2}(d_N) & \dots & f_{Ip}(d_N) \end{bmatrix}$$

where p is the dimension of these features, f_{Ik} indicates the function of extracting property, $k = 1, \dots, p$.

4) USER FEATURES

Netizens whose account information is more detailed and authoritative are more cautious when posting news. Users, who are not authenticated and have a short account registration time, have fewer concerns and lower reliability when publishing fake news [37]. Therefore, we detect the credibility of users through users' information, influence, and behavior. The user features can be indicated as matrix A_4, as the following:

$$A_4 = \begin{bmatrix} f_{U1}(d_1) & f_{U2}(d_1) & \dots & f_{Uq}(d_1) \\ f_{U1}(d_2) & f_{U2}(d_2) & \dots & f_{Uq}(d_2) \\ \dots & \dots & \dots & \dots \\ f_{U1}(d_N) & f_{U2}(d_N) & \dots & f_{Uq}(d_N) \end{bmatrix}$$

where q is the dimension of these features, f_{Uv} indicates the function of extracting property, $v = 1, \dots, q$.

In summary, the details of these four features as seen in Table 1.

When fusing these four kinds of features, this paper uses the splicing method. Here, let text content features $X_C = [f_{C1}, f_{C2}, \dots, f_{Cm}]$, propagation features $X_P = [f_{P1}, f_{P2}, \dots, f_{Pn}]$, user features $X_U = [f_{U1}, f_{U2}, \dots, f_{Uq}]$, image features $X_I = [f_{I1}, f_{I2}, \dots, f_{Ip}]$, then the vector X after fusing the four kinds of features can be described as follows, where $d = m + n + p + q$.

$$X = concat[X_C; X_P; X_I; X_U] = [F_1, F_2, \dots, F_d] \quad (7)$$

B. THE MODEL BASED ON AUTOENCODER

As shown in Figure 2, after integrating various features of news in social networks, this paper proposes the model based on autoencoder to analyze the hidden information and the internal relations, and then realizes unsupervised fake news detection. This section is the introduction to the model.

TABLE 1. Details of extraction features of fake news detection.

Feature types	Feature names
Text content features	number of positive sentiment Words, number of negative sentiment Words, number of slangs, contains question mark, number of question mark, contains first-person pronouns, contains second-person pronouns, contains third-person pronouns, has please, contains happy emotion, contains sad emotion, number of exclamation mark, length of text, has external link, number of uppercase chars, number of URLs, number of words, number of hashtags, contains exclamation mark, has colon, number of nouns, number of mentions
Propagation features	number of retweet, number of comment, number of praise
Image features	is tampered
User features	number of media content, has existing location, tweet ratio, number of tweets, is verified, number of friends, has bio description, has header image, has location, has profile IMG, follower-friend ratio, account age, has URL, number of followers, times listed, number of favorites

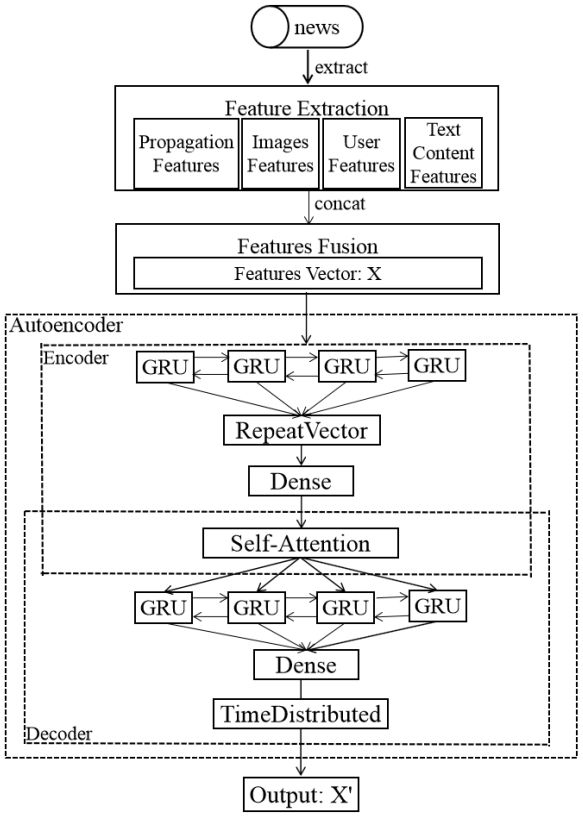


FIGURE 2. Autoencoder-based method for unsupervised fake news detection.

1) AUTOENCODER

Autoencoder is a data compression algorithm, and its Encoder and Decoder are realized by neural network. The proposed method (UFNDA) based on AE is an error reconstruction anomaly detection method, and news with high reconstruction is regarded as fake news. Figure 3 is the autoencoder structure of UFNDA.

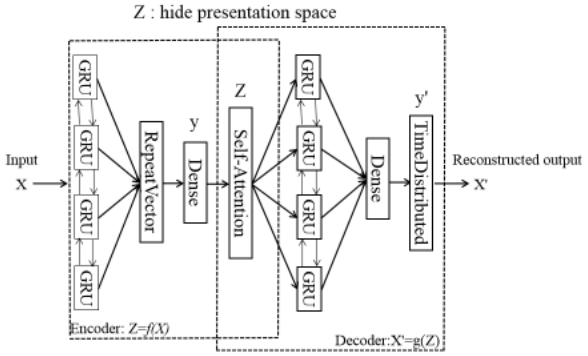


FIGURE 3. The autoencoder structure of UFNDA.

In this structure, the Dense layer is to transform the dimension of the feature space. For making each time step to obtain the same input but different hidden states, the RepeatVector layer is combined with the encoder of autoencoder. To reduce the dimensionality of the feature space on each time step, the TimeDistributed layer is combined with the decoder layer of autoencoder.

a: ENCODER

The features vector X, as input data, is encoded into the latent space Z by function f, and Z is as follows:

$$Z = f(X) = \sigma(self - attention(Dense(RepeatVector(Bi - GRU(X))))) \quad (8)$$

b: DECODER

Decode Z into X' through function g, and X' is as follows:

$$X' = g(Z) = \sigma'(TimeDistributed(Dense(Bi - GRU(Z)))) \quad (9)$$

Therefore, $g(f(X)) = X'$ can describe the entire AE, which makes the X' as close to the X as possible and trains our method by minimizing the reconstruction error between the X and the X'. The Root Mean Square Error (RMSE) is used to measure the error, as the following equation, where N represents the number of news data.

$$L(X, X') = ||X' - X|| = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(X' - X)^2} \quad (10)$$

2) BI-GRU

Gated Recurrent Unit (GRU) [38] is a variant of recurrent neural network (RNN), which can overcome the gradient disappearance and explosion problems in RNN, and solve long-term and short-term dependence. To obtain the hidden information between the features, we combine Bi-GRU layer that can capture the two-way relationship with autoencoder, as seen in Figure 2. Each Bi-GRU layer includes a forward $\overrightarrow{GRU}$ and a backward $\overleftarrow{GRU}$, and each GRU includes an input layer and a self-connected hidden layer, as seen in Figure 4.

The forward $\overrightarrow{GRU}$ takes the features vector $\overrightarrow{X}$ as input, as seen in Figure 5. At t time, the feature vector x_t is sent to

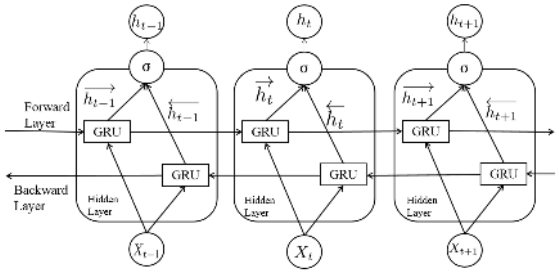


FIGURE 4. The Bi-GRU structure.

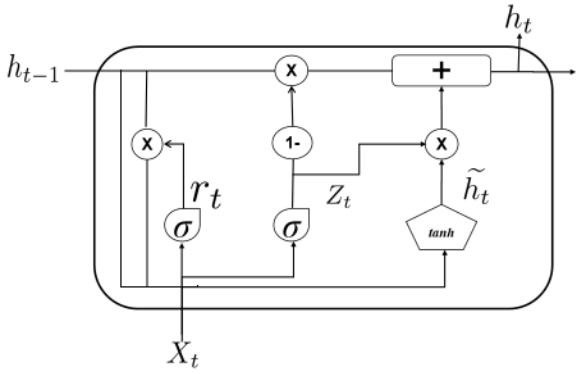


FIGURE 5. The forward GRU structure.

the input layer of $\overrightarrow{GRU}$, combines with the forward hidden state $\overrightarrow{h_{t-1}}$, and then perform a series of linear operations and activation operations. Each GRU has an update gate Z_t and a reset gate r_t. The current state and the amount of historical information retained in the gate are controlled by Z_t, and r_t determines the strength of ignoring irrelevant information sequences to ensure that important information sequences are delivered to the moment. The process of $\overrightarrow{GRU}$ is as follows:

$$Z_t = \sigma(W_Z[\overrightarrow{h_{t-1}}, x_t]) \quad (11)$$
$$r_t = \sigma(W_r[\overrightarrow{h_{t-1}}, x_t) \quad (12)$$
$$\widetilde{h_t} = tanh(W[r_t\overrightarrow{h_{t-1}}, x_t]) \quad (13)$$
$$\overrightarrow{h_t} = (1 - Z_t)\overrightarrow{h_{t-1}} + Z_t\widetilde{h_t} \quad (14)$$

where W_z and W_r are weight matrices, σ represents the sigmoid activation function.

The backward $\overleftarrow{GRU}$ takes the features vector $\overleftarrow{X}$ as input. At t time, execute a series of calculations which is similar to the forward $\overrightarrow{GRU}$, and obtain the backward $\overleftarrow{h_t}$ which combines with the forward $\overrightarrow{h_t}$ to form a hidden state with bidirectional information, where d is the dimension of X, then$ht = [\overrightarrow{h_t}, \overleftarrow{h_t}]$. The hidden layer of Bi-GRU is as follows:

$$h = [h_1, h_2, \ldots, h_t] = [[\overrightarrow{h_1}, \overleftarrow{h_1}], [\overrightarrow{h_2}, \overleftarrow{h_2}], .., [\overrightarrow{h_t}, \overleftarrow{h_t}]] \quad (15)$$

To better enhance features, reduce errors, and speed up convergence, RELU is used as the activation function, then

$y_t = \text{ReLU}(h_t) = \text{Bi-GRU}(X_t)$, and the final output of the Bi-GRU layer $y = [y_1, y_2, \ldots, y_t]$.

3) SELF-ATTENTION

Self-attention is a new breakthrough in attention technology, which connects different positions of a single sequence to calculate an attention mechanism [39].

To obtain the internal connection between each feature, we combine the self-attention layer with autoencoder, as seen in Figure 2. Self-attention is a special case of the general attention. It is constructed by Query (Q), Key (K), and Value (V), and to extract the internal information by itself. Query can be regarded as a giver, Key can be regarded as a receiver, and Value can be regarded as an information extractor, where $Q = K = V$. Each unit in one sequence and all units in sequences perform attention operations.

Figure 6 is the calculation process of the self-attention layer. The input of this layer comes from the output of the Bi-GRU layer $y = y_1, y_2, \ldots, y_t$. In this calculation, $K = W^k y^i, Q = W^q y^i, V = W^v y^i$, the relationships of y^i are first constructed by the Q and the K, then all the internal relationships of y^i are summarized by the V, finally output Z is obtained, where $\alpha_{i,j} = q_i \cdot k_j / \sqrt{e}$, e is the dimension of q and k, $\tilde{a}_{i,j} = Softmax(\alpha_{i,j})$, $Z_i = \sum \tilde{\alpha}_{i,j} \cdot v^j$, $Z_t = Self - Attention(y_t)$.

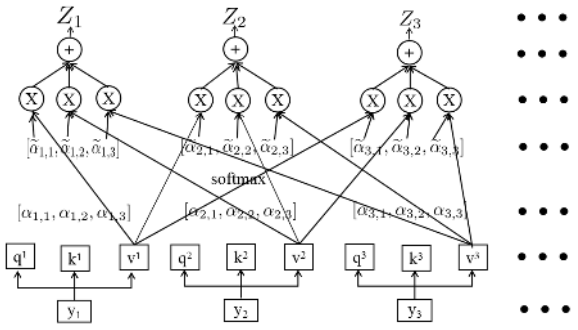


FIGURE 6. Calculation process of the self-attention.

C. FAKE NEWS DETECTION METHOD

Algorithm 1 is the fake news detection algorithm in UFNDA, which fuses the extracted features, then analyzes the hidden information and internal relations between the features. It is based on the autoencoder of self-attention method, and obtains the residual value of the news to be detected. It is a piece of fake news if its residual value is larger than the threshold, otherwise, it is a piece of real news.

Algorithm 1 Fake News Detection Algorithm in UFNDA

INPUT: The number of fake news N_{error}
OUTPUT: reconstruction error $\| x' - x \|$

```
Datasets ← get news text content from social networks APIs
Datasets ← get news propagation info from social networks APIs
Datasets ← get news image info from online fact-checking resources
Datasets ← get user info from social networks APIs
X_C ← extract features using Datasets
X_P, X_I, X_U ← extract features using Datasets
X ← fuse features using X_C, X_P, X_I, X_U
X_train ← get real dataset using X
x^(i) i = 1, 2, 3, ..., n ← get test dataset using X
φ, θ ← train the method using X_train
for i = 1 to n do
    reconstruction error(i) = ‖ g_θ(f_φ(x^(i))) − x^(i) ‖
end for
re_sorted = sort(reconstruction error, reverse = True)
α ← get threshold from re_sorted where index = N_error − 1
for i = 1 to n do
    if reconstruction error(i) > α then
        x^(i) is a piece of fake news
    else
        x^(i) is a piece of real news
    end if
end for
```

IV. EXPERIMENTS AND ANALYSIS

For conducting the experiments and evaluating the effectiveness of the proposed method UFNDA, this paper makes use of a Twitter dataset and a Weibo dataset.

A. DATASETS

The Twitter dataset published by Verifying Multimedia Use at MediaEval 2016 [40]. This dataset comes from 17 events and contains 17,857 data from 15,821 users, which includes 7,244 real data and 10613 fake data.

The Weibo dataset gains from Sina Weibo - a Chinese biggest social network platform. This platform provides an online fact-checking function to censor any suspect content violated social network behaviors and rules. A review committee will examine and verify the reliability of these suspect content without pay. This paper collects data set containing 5376 real news and 562 fake news from this platform.

Detecting fake news as anomalies requires an unbalanced dataset. The statistics of the two datasets shown as Table 2.

TABLE 2. The detail information of the two datasets.

Statistics		Twitter	Weibo
Train	Real	6339	4779
Validation	Real	905	597
Test	Real	6000	4500
	Fake	750	562

B. EVALUATION METRICS

AUC, Macro-F1, Micro-F1, and Precision are used to evaluate the performance of UFNDA, and their concepts are described as follows.

1) AUC

AUC is the area under the ROC curve. ROC (Receiver Operating Characteristics) curve is a probability curve that shows the effect of the classification model under all classification thresholds. The true rate (TPR) and the false positive rate (FPR) can be used to described as ROC which defined as follows:

$$TPR = \frac{TP}{TP + FN} \tag{16}$$

$$FPR = \frac{FP}{FP + TN} \tag{17}$$

The value range of AUC is 0 1. The closer the AUC is to 1, the better the performance of the method in detecting fake news; the closer to 0, the worse the performance of the method.

2) MACRO-F1

Macro-F1 is to first calculate the metric values of each category, and then find the arithmetic average of all categories. It is defined as follows:

$$Macro\_F1 = \frac{2 \cdot Macro\_Precision \cdot Macro\_Recall}{Macro\_Precision + Macro\_Recall} \tag{18}$$

3) MICRO-F1

Micro-F1 is to calculate each instance in dataset and ignore categories, build a global confusion matrix, and then calculate the related indexes. It is defined as follows:

$$Micro\_F1 = \frac{2 \cdot Micro\_Precision \cdot Micro\_Recall}{Micro\_Precision + Micro\_Recall} \tag{19}$$

4) PRECISION

Precision is calculated as the ratio of the number of samples properly classified as positive to the sum of samples classified as positive. It is defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{20}$$

C. RESULTS ANALYSIS

The experiments include two parts:

1) Verify the effectiveness of the proposed method UFNDA. Compare the performance of UFNDA to other existing methods, including fake news detection and early fake news detection.

2) Evaluate the effectiveness of features fusion. Compare the performance of UFNDA to the performance of UFNDA based on other feature combinations.

1) UFNDA VALIDATION

Isolation Forest (ISF) [41], One-Class SVM (OCSVM) [42], autoencoder (AE), and Variational autoencoder (VAE) [43] are selected for the comparative experiments to evaluate the performance of the different unsupervised learning classifiers to fake news detection. Table 3, Figure 7, Table 4, and Figure 8 are the comparison of experimental results.

TABLE 3. The comparison of experimental results on Twitter.

Model	AUC	Macro-F1	Micro-F1
OCSVM	0.5925	0.4892	0.7982
ISF	0.5772	0.5073	0.8053
VAE	0.5045	0.4953	0.8006
AE	0.5968	0.4953	0.8030
UFNDA	**0.6489**	**0.5088**	**0.8059**

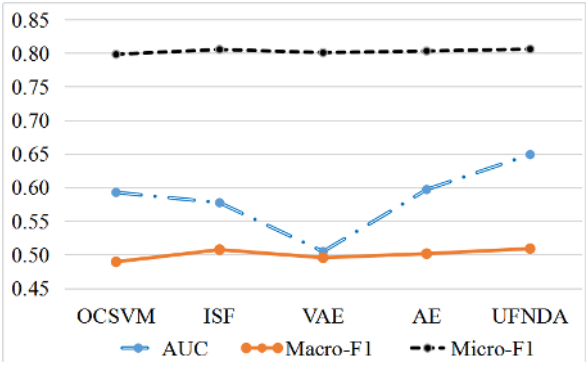


FIGURE 7. The comparison of experimental results on Twitter.

TABLE 4. The comparison of experimental results on Weibo.

Model	AUC	Macro-F1	Micro-F1
OCSVM	0.6824	0.5657	0.8285
ISF	0.6911	0.5837	0.8356
VAE	0.6676	0.5677	0.8293
AE	0.6914	0.5867	0.8368
UFNDA	**0.7017**	**0.6097**	**0.8459**

It can be seen from Table 3 and Figure 7 that the proposed method UFNDA performs best in these three evaluation metrics. When only looking at the AUC, AE performs better than OCSVM, ISF, and VAE. The UFNDA is improved based on AE. Experimental results show that AUC of the improved model (UFNDA) is 5.21% higher than the original model (AE). This is because the Bi-GRU and self-attention layers are added to the proposed method UFNDA, which is more capable of capturing and organizing hidden information than the AE with only the fully connected layers. On Macro-F1 and Micro-F1, AE performance is not as good as ISF performance, but the improved method UFNDA performs better than other four methods. Because OCSVM and ISF detect singularities based on spatial distance, while VAE and AE are based on neural networks, but the improved method UFNDA can obtain more hidden information and improve the model's efficiency. These conclusions can be seen from Table 4 and Figure 8.

The value of AUC exceeds 0.5 indicates these methods are all effective but UFNDA has better ability of classification. Macro-F1 treats each category equally and takes the average of all categories. However, Micro-F1 accords the contributions of different categories to calculate the average. What's more, this paper uses unbalanced dataset. Therefore, the value of Micro-F1 is larger and more reliable than that of Macro-F1.

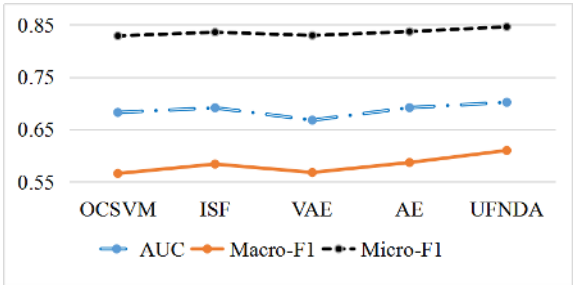


FIGURE 8. The comparison of experimental results on Weibo.

This paper designed an experiment to obtain the precision of the real news and the fake news, which also proves the effectiveness of the proposed method UFNDA to a certain extent, as shown in Table 5.

TABLE 5. The results of fake news detection in precision on the two datasets.

Datasets	Model	Precision	
		Real News	Fake News
Twitter	OCSVM	0.8870	0.0960
	ISF	0.8913	0.1307
	VAE	0.8878	0.1027
	AE	0.8892	0.1133
	UFNDA	**0.8983**	**0.1357**
Weibo	OCSVM	0.9036	0.2278
	ISF	0.9087	0.2687
	VAE	0.9040	0.2313
	AE	0.9082	0.2651
	UFNDA	**0.9106**	**0.2804**

In addition, the experiment on early fake news detection is also conducted, and the results as seen in Table 6, Figure 9, Table 7, and Figure 10.

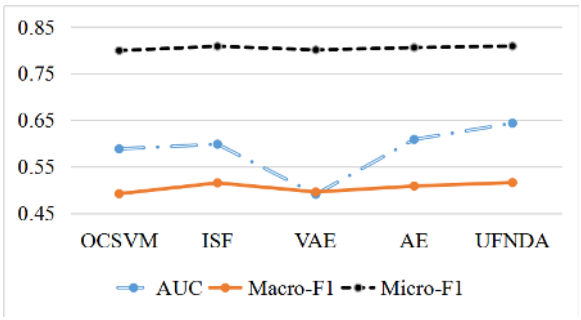


FIGURE 9. The comparison of experimental results of early fake news detection on Twitter.

From Table 6, Figure 9, Table 7, and Figure 10, we can see that UFNDA also has a better performance than other four methods on the two datasets, which indicates that the proposed method is also effective for early fake news detection.

2) FEATURES FUSION VALIDATION

The experiments of the proposed method UFNDA based on the feature combinations, including text content features (C), propagation features (P), image features (I) and user features (U), are also designed. The experimental results as seen in Table 8, Figure 11, Table 9, and Figure 12.

We can observe from Table 8 and Figure 11 that text content features of news and image features of news, on Twitter dataset, are important, particularly image features. The UFNDA performs not the best which indicates that a simple splicing fusion does not suite for the data from the social platform of Twitter.

TABLE 6. The comparison of experimental results of early fake news detection on Twitter.

Model	AUC	Macro-F1	Micro-F1
OCSVM	0.5879	0.4915	0.7991
ISF	0.5980	0.5148	0.8083
VAE	0.4902	0.4953	0.8006
AE	0.6082	0.5080	0.8056
UFNDA	**0.6431**	**0.5155**	**0.8086**

TABLE 7. The comparison of experimental results for early fake news detection on Weibo.

Model	AUC	Macro-F1	Micro-F1
OCSVM	0.5664	0.5056	0.8048
ISF	0.6057	0.5326	0.8155
VAE	0.5436	0.5186	0.8100
AE	0.5531	0.5086	0.8060
UFNDA	**0.6839**	**0.5556**	**0.8246**

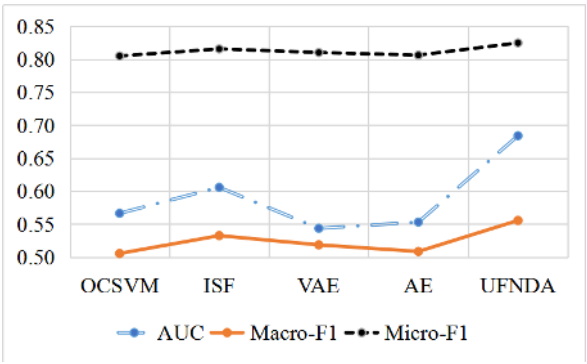


FIGURE 10. The comparison of experimental results for early fake news detection on Weibo.

TABLE 8. The experimental results of UFNDA based on feature combinations on Twitter.

Method	AUC	Macro-F1	Micro-F1
UFNDA_C	0.6233	**0.5260**	0.8127
UFNDA_U	0.5284	0.4825	0.7956
UFNDA_I	**0.8128**	0.4720	**0.8890**
UFNDA_P	0.5515	0.4756	0.8413
UFNDA_CU	0.5752	0.5035	0.8039
UFNDA_CP	0.6046	0.5230	0.8116
UFNDA_CI	0.6378	0.5238	0.8110
UFNDA_UI	0.6742	0.4855	0.7967
UFNDA_UP	0.5235	0.4803	0.7947
UFNDA_IP	0.8073	0.4845	0.8796
UFNDA_CUP	0.6004	0.5125	0.8074
UFNDA_CUI	0.6431	0.5155	0.8086
UFNDA_CIP	0.6644	0.5515	0.8228
UFNDA_UPI	0.6566	0.4758	0.7929
UFNDA_CUPI(UFNDA)	0.6489	0.5088	0.8059

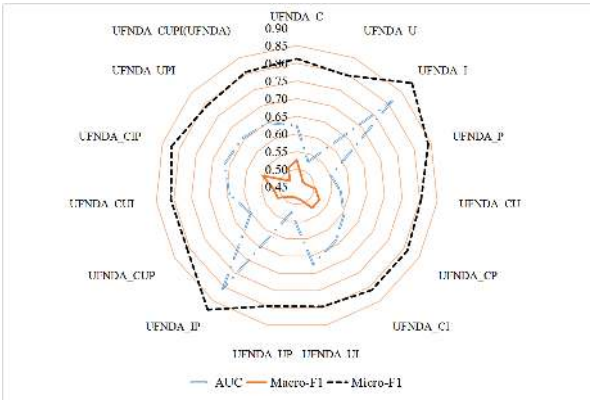


FIGURE 11. The experimental results of UFNDA based on feature combinations on Twitter.

TABLE 9. The experimental results of UFNDA based on feature combinations on Weibo.

Method	AUC	Macro-F1	Micro-F1
UFNDA_C	0.4862	0.5356	0.8167
UFNDA_U	0.8173	0.6200	0.8572
UFNDA_I	0.7591	0.5506	0.8226
UFNDA_P	0.7780	**0.7538**	**0.9028**
UFNDA_CU	0.6091	0.5506	0.8226
UFNDA_CP	0.4901	0.5356	0.8167
UFNDA_CI	0.6162	0.5587	0.8258
UFNDA_UI	0.8235	0.6057	0.8443
UFNDA_UP	0.8250	0.6357	0.8562
UFNDA_IP	0.7628	0.5597	0.8262
UFNDA_CUP	0.5958	0.5396	0.8183
UFNDA_CUI	0.6839	0.5556	0.8246
UFNDA_CIP	0.5789	0.5356	0.8167
UFNDA_UPI	**0.8305**	0.6277	0.8530
UFNDA_CUPI(UFNDA)	0.7017	0.6097	0.8459

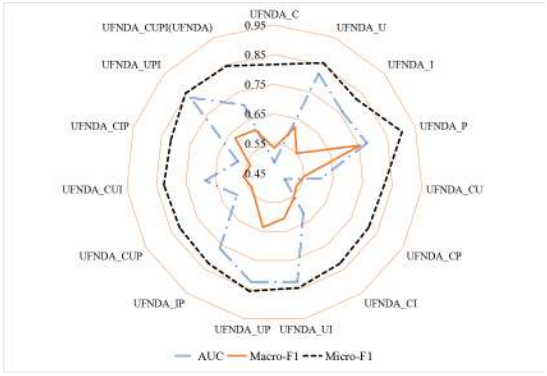


FIGURE 12. The experimental results of UFNDA based on feature combinations on Weibo.

Table 9 and Figure 12 display that propagation features of news, the features of user who publish news, and image features of news, on Weibo dataset, are important, particularly propagation features. The UFNDA based on these three kinds of features have better results than UFNDA based on a single feature, which indicates these three are more suitable for splicing fusion on the data from the social platform of Weibo.

In summary, the UFNDA based on these features perform better than the UFNDA based on one of these features, but that's not always the case. This conclusion also illuminates that simple splicing fusion may not coordinate data from different social platforms with different features.

V. CONCLUSION

This paper proposes a method based on autoencoder to solve the problem of unsupervised fake news detection. First, extract and fuse Twitter's text content features, propagation features, image features, and user features in social networks, then use fake news as anomalous data and analyze it with the proposed method UFNDA, and finally classify the test data by reconstructing errors. Experimental results show that UFNDA is superior to several methods in unsupervised fake news detection.

Although UFNDA performs well, there are still some improvements. In social networks, there are many features, such as comments, dissemination of news, and videos, are important for fake news detection. In addition, current news are not completely fake or completely true, and more detailed classification needs to be considered. These are the next research plans of our work.

REFERENCES

[1] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," Nat. Bur. Econ. Res., Cambridge, MA, USA, Tech. Rep. w23089, 2017.
[2] P. Meel and D. K. Vishwakarma, "Fake news, rumor, information pollution in social media and Web: A contemporary survey of state-of-the-arts, challenges and opportunities," *Expert Syst. Appl.*, vol. 153, Sep. 2020, Art. no. 112986.
[3] S. Ghosh and C. Shah, "Towards automatic fake news classification," *Proc. Assoc. Inf. Sci. Technol.*, vol. 55, no. 1, pp. 805–807, Jan. 2018.
[4] Y. Liu and Y.-F. Wu, "Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks," in *Proc. 32th AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.
[5] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explor. Newslett.*, vol. 19, no. 1, pp. 22–36, Sep. 2017.
[6] X. Chen, F. Zhou, K. Zhang, G. Trajcevski, T. Zhong, and F. Zhang, "Information diffusion prediction via recurrent cascades convolution," in *Proc. IEEE 35th Int. Conf. Data Eng. (ICDE)*, Macao, China, Apr. 2019, pp. 770–781, doi: 10.1109/ICDE.2019.00074.
[7] X. Zhou, R. Zafarani, K. Shu, and H. Liu, "Fake news: Fundamental theories, detection strategies and challenges," in *Proc. 12th ACM Int. Conf. Web Search Data Mining*, Melbourne, VIC, Australia, 2019, pp. 836–837.
[8] F. A. Ozbay and B. Alatas, "Fake news detection within online social media using supervised artificial intelligence algorithms," *Phys. A, Stat. Mech. Appl.*, vol. 540, Feb. 2020, Art. no. 123174.
[9] J. Z. Pan, S. Pavlova, C. Li, N. Li, Y. Li, and J. Liu, "Content based fake news detection using knowledge graphs," in *Proc. Int. Semantic Web Conf.*, 2018, pp. 669–683.
[10] R. A. Stein, P. A. Jaques, and J. F. Valiati, "An analysis of hierarchical text classification using word embeddings," *Inf. Sci.*, vol. 471, pp. 216–232, Jan. 2019.
[11] H. Seyedmehdi and E. Papalexakis, "Unsupervised content-based identification of fake news articles with tensor decomposition ensembles," in *Proc. Misinf. Misbehav. Mining Web Workshop Held Conjunct WSDM (MIS2)*, Los Angeles, CA, USA, 2018, pp. 1–8.
[12] S. Yang, K. Shu, S. Wang, R. Gu, F. Wu, and H. Liu, "Unsupervised fake news detection on social media: A generative approach," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 01, pp. 5644–5651.
[13] J. Gaglani, Y. Gandhi, S. Gogate, and A. Halbe, "Unsupervised WhatsApp fake news detection using semantic search," in *Proc. 4th Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, Madurai, India, May 2020, pp. 285–289, doi: 10.1109/ICICCS48265.2020.9120902.
[14] S. C. R. Gangireddy, D. P, C. Long, and T. Chakraborty, "Unsupervised fake news detection: A graph-based approach," in *Proc. 31st ACM Conf. Hypertext Social Media*, Jul. 2020, pp. 75–83.

[15] M. Farajtabar, J. Yang, X. Ye, H. Xu, R. Trivedi, E. Khalil, S. Li, L. Song, and H. Zha, "Fake news mitigation via point process based intervention," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 1097–1106.

[16] N. Renström, P. Bangalore, and E. Highcock, "System-wide anomaly detection in wind turbines using deep autoencoders," *Renew. Energy*, vol. 157, pp. 647–659, Sep. 2020.

[17] Z. Luo, Y. Xiong, and R. Zuo, "Recognition of geochemical anomalies using a deep variational autoencoder network," *Appl. Geochem.*, vol. 122, Nov. 2020, Art. no. 104710.

[18] Q. Feng, Z. Dou, C. Li, and G. Si, "Anomaly detection of spectrum in wireless communication via deep autoencoder," in *Advances in Computer Science and Ubiquitous Computing* (Lecture Notes in Electrical Engineering), vol. 421, J. Park, Y. Pan, G. Yi, and V. Loia, Eds. Singapore: Springer, 2017.

[19] R. Ferri, C. Babiloni, V. Karami, A. I. Triggiani, F. Carducci, G. Noce, R. Lizio, M. T. Pascarelli, A. Soricelli, F. Amenta, A. Bozzao, A. Romano, F. Giubilei, C. D. Percio, F. Stocchi, G. B. Frisoni, F. Nobili, L. Patané, and P. Arena, "Stacked autoencoders as new models for an accurate Alzheimer's disease classification support using resting-state EEG and MRI measurements," *Clin. Neurophysiol.*, vol. 132, no. 1, pp. 232–245, Jan. 2021.

[20] M. Porumb, C. Griffen, J. Hattersley, and L. Pecchia, "Nocturnal low glucose detection in healthy elderly from one-lead ECG using convolutional denoising autoencoders," *Biomed. Signal Process. Control*, vol. 62, Sep. 2020, Art. no. 102054.

[21] H. Akrami, A. A. Joshi, J. Li, S. Aydore, and R. M. Leahy, "Brain lesion detection using a robust variational autoencoder and transfer learning," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Iowa City, IA, USA, Apr. 2020, pp. 786–790, doi: 10.1109/ISBI45749.2020.9098405.

[22] S. Koppers, E. Coussoux, S. Romanzetti, K. Reetz, and D. Merhof, "Sodium image denoising based on a convolutional denoising autoencoder," in *Bildverarbeitung für die Medizin*. Wiesbaden, Germany: Springer-Verlag, 2019, doi: 10.1007/978-3-658-25326-4_23.

[23] D. Oszutowska-Mazurek, P. Mazurek, and O. Knap, "Stacked autoencoder for segmentation of bone marrow histological images," in *Artificial Intelligence and Algorithms in Intelligent Systems* (Advances in Intelligent Systems and Computing), vol. 764, R. Silhavy, Ed. Cham, Switzerland: Springer, 2019.

[24] S. A. Deepthi, E. S. Rao, and M. N. G. Prasad, "Image compression in wireless sensor networks using autoencoder and RBM method," in *Innovations in Electronics and Communication Engineering* (Lecture Notes in Networks and Systems), vol. 65, H. Saini, R. Singh, G. Kumar, G. Rather, and K. Santhi, Eds. Singapore: Springer, 2019.

[25] M. Mendoza, B. Poblete, and C. Castillo, "Twitter under crisis: Can we trust what we RT?" in *Proc. 1st Workshop Social Media Anal.*, Washington, DC, USA, 2010, pp. 71–79.

[26] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *Proc. IEEE 13th Int. Conf. Data Mining (ICDM)*, Dallas, TX, USA, Dec. 2013, pp. 1103–1108, doi: 10.1109/ICDM.2013.61.

[27] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel visual and statistical image features for microblogs news verification," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 598–608, Mar. 2017, doi: 10.1109/TMM.2016.2617078.

[28] M. D. Vicario, W. Quattrociocchi, A. Scala, and F. Zollo, "Polarization and fake news: Early warning of potential misinformation targets," *ACM Trans. Web*, vol. 13, no. 2, pp. 1–22, Apr. 2019.

[29] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," in *Proc. Int. Conf. Intell., Secur., Dependable Syst. Distrib. Cloud Environ.*, 2017, pp. 127–138.

[30] J. Fairbanks, N. Fitch, N. Knauf, and E. Briscoe, "Credibility assessment in the news: Do we need to read?" in *Proc. MIS2*, 2018, pp. 799–800.

[31] K. Shu, S. Wang, and H. Liu, "Beyond news contents: The role of social context for fake news detection," in *Proc. 12th ACM Int. Conf. Web Search Data Mining*, Jan. 2019, pp. 312–320.

[32] Y. Yang, L. Zheng, J. Zhang, Q. Cui, Z. Li, and P. S. Yu, "TI-CNN: Convolutional neural networks for fake news detection," 2018, *arXiv:1806.00749*. [Online]. Available: http://arxiv.org/abs/1806.00749

[33] S. Esmaeilzadeh, G. X. Peh, and A. Xu, "Neural abstractive text summarization and fake news detection," 2019, *arXiv:1904.00788*. [Online]. Available: http://arxiv.org/abs/1904.00788

[34] R. K. Kaliyar, A. Goswami, P. Narang, and S. Sinha, "FNDNet—A deep convolutional neural network for fake news detection," *Cogn. Syst. Res.*, vol. 61, pp. 32–44, Jun. 2020.

[35] A. Uppal, V. Sachdeva, and S. Sharma, "Fake news detection using discourse segment structure analysis," in *Proc. 10th Int. Conf. Cloud Comput., Data Sci. Eng. (Confluence)*, Noida, India, Jan. 2020, pp. 751–756, doi: 10.1109/Confluence47617.2020.9058106.

[36] W. Antoun, F. Baly, R. Achour, A. Hussein, and H. Hajj, "State of the art models for fake news detection tasks," in *Proc. IEEE Int. Conf. Informat., IoT, Enabling Technol. (ICIoT)*, Doha, Qatar, Feb. 2020, pp. 519–524, doi: 10.1109/ICIoT48696.2020.9089487.

[37] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Inf. Process. Manage.*, vol. 57, no. 2, Mar. 2020, Art. no. 102025.

[38] K. Cho, B. V. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1–15. [Online]. Available: https://arxiv.org/abs/1406.1078

[39] 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. Neural Inf. Process. Syst.* Cambridge, MA, USA: MIT Press, 2017, pp. 5998–6008.

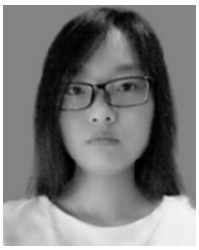
[40] C. Boididou, S. Papadopoulos, D. Dang-Nguyen, G. Boato, M. Riegler, S. E. Middleton, A. Petlund, and Y. Kompatsiaris, "Verifying multimedia use at mediaeval 2016," in *Proc. MediaEval Workshop*, 2016, pp. 1–3.

[41] F. T. Liu, K. M. Ting, and Z. Zhou, "Isolation-based anomaly detection," *ACM Trans. Knowl. Discovery Data*, vol. 6, no. 1, pp. 1–39, Mar. 2012.

[42] S. S. Khan and M. G. Madden, "A survey of recent trends in one class classification," in *Proc. Irish Conf. Artif. Intell. Cogn. Sci.* Berlin, Germany: Springer, 2009, pp. 188–197.

[43] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 1–14.

DUN LI received the Ph.D. degree from the Beijing Institute of Technology, in 2007. She is currently an Associate Professor with the School of Information Engineering, Zhengzhou University. Her main research interests include intelligent information processing and social networks.



HAIMEI GUO received the B.E. degree in computer science and technology from Zhengzhou University, Zhengzhou, China, where she is currently pursuing the M.S. degree in computer science and technology. Her research interests include social networks, NLP, data analysis, and machine learning.

ZHENFEI WANG received the Ph.D. degree from the Huazhong University of Science and Technology, China, in 2006. He is currently a Professor with the School of Information, Zhengzhou University, China. His current research interests include data mining, social networks, and intelligent medical treatment. He is a member of the China Computer Society.

ZHIYUN ZHENG received the Ph.D. degree from the Beijing Institute of Technology, in 2005. She is currently a Professor with the School of Information Engineering, Zhengzhou University. Her current research interests include cloud computing and intelligent information processing.