Information Processing and Management 58 (2021) 102610

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Detecting fake news by exploring the consistency of multimodal data

Junxiao Xue [a], Yabo Wang [a,*], Yichen Tian [a], Yafei Li [b], Lei Shi [a], Lin Wei [a]

[a] School of Software, Zhengzhou University, 450002, China
[b] School of Information Engineering, Zhengzhou University, 450001, China

ARTICLE INFO

Keywords:
Fake news detect
Multimodal
Social media
Neural network
Tampering

ABSTRACT

During the outbreak of the new Coronavirus (2019-nCoV) in 2020, the spread of fake news has caused serious social panic. Fake news often uses multimedia information such as text and image to mislead readers, spreading and expanding its influence. One of the most important problems in fake news detection based on multimodal data is to extract the general features as well as to fuse the intrinsic characteristics of the fake news, such as mismatch of image and text and image tampering. This paper proposes a Multimodal Consistency Neural Network (MCNN) that considers the consistency of multimodal data and captures the overall characteristics of social media information. Our method consists of five subnetworks: the text feature extraction module, the visual semantic feature extraction module, the visual tampering feature extraction module, the similarity measurement module, and the multimodal fusion module. The text feature extraction module and the visual semantic feature extraction module are responsible for extracting the semantic features of text and vision and mapping them to the same space for a common representation of cross-modal features. The visual tampering feature extraction module is responsible for extracting visual physical and tamper features. The similarity measurement module can directly measure the similarity of multimodal data for the problem of mismatching of image and text. We assess the constructed method in terms of four datasets commonly used for fake news detection. The accuracy of the detection is improved clearly compared to the best available methods.

1. Introduction

Today's world is the era of self media and everyone can produce content and contribute to public opinion. Videos and images can narrate and engage readers better than text-only content. Unfortunately, these are also used by fake news. Fake news uses fictional or even fake pictures to mislead readers and spread quickly (Allcott & Gentzkow, 2017; Rubin, Conroy, Chen, & Cornwell, 2016). The spread of fake news may cause massive negative effects, sometimes affecting or manipulating major public events. The 2016 United States presidential election showed the concerns of the public about the fake news influencing the citizens' impression on candidates. During the new Coronavirus (2019-nCoV), fake epidemic news spreads throughout the Internet, causing a great psychological panic among citizens. At the recent Munich Security Conference, Google released a white paper that also highlighted the need to open the Internet to combat fake news. Thus, there is an indication that research on fake news is urgent. Eliminating fake news is of great necessity for perfecting the quality of the information network ecosystem and maintaining social stability (Castillo, Mendoza, & Poblete, 2011; Qian, Gong, Sharma, & Liu, 2018; Ruchansky, Seo, & Liu, 2017; Sunstein, 2009).

* Corresponding author.
E-mail address: wangyb@stu.zzu.edu.cn (Y. Wang).

https://doi.org/10.1016/j.ipm.2021.102610
Received 29 October 2020; Received in revised form 14 April 2021; Accepted 17 April 2021
Available online 3 May 2021

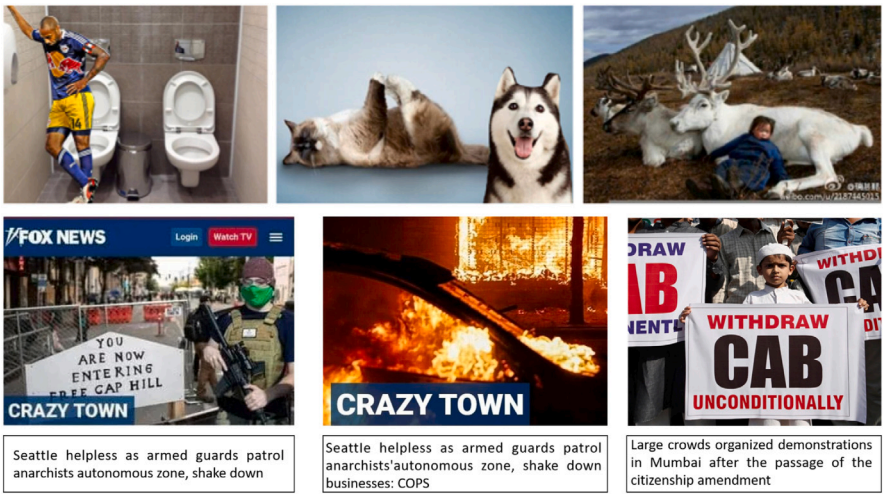

Fig. 1. Some fake news with its image.

As an intentional and verifiable fake news article, fake news content usually contains textual and visual information (Jin, Cao, Zhang, Zhou and Tian, 2017; Qi, Cao, Yang, Guo, & Li, 2019). As shown in Fig. 1, fake news publishers often use text and image that fabricate or misrepresent facts to cater to readers' psychology which can attract and mislead readers for rapid dissemination. Generally, the topics which focus on social hot spots or controversies have intense textual descriptions on their emotional expression and visual impact on images. Fake news could come under various modes of data like texts, pictures and videos. It is a collection of multimedia which means it is hard to detect fake news from single-modal data.

Each modal data can feed rumors in different degrees. Meanwhile, fake news often uses the content like pictures or texts with highly sentiment orientation to spread quickly. However, there is not much research on how these modes affect the news. Therefore, it is necessary to integrate the multimodal features of news to detect fake news.

The present methods of fake news detection are based on either single modal data or merge two types of data (Ma et al., 2016; Ma, Gao, & Wong, 2019). These approaches ignore the effective modeling of various modalities and the similarity between multimodal data (Wang et al., 2018; Yang et al., 2018; Zhou, Wu, & Zafarani, 2020). As a result, it is impossible to deeply dig into the inherent characteristics of fake news (such as image tampering, inconsistent images, etc.). What makes the text of fake news different from visual information is that some fake news (or news with low credibility) use theatrical, comical and attractive images to catch the publics' eyes, resulting in the textual content of the news being removed from the actual content. Ma et al. (2016) first included social network multimodal content that solves the problems of fake news detection by using deep neural networks. Wang et al. (2018) put forward an end-to-end event adversarial neural network based on multimodal features to detect emerging fake news events. However, these works are mainly instructive and ignore the effective modeling of visual content. The visual features used by them are mainly obtained by pre-trained convolutional neural network like VGG19, which is hard to show the intrinsic features of a fake news image due to lack of task-relevant information (Antol et al., 2015; Lin, He, Tang, & Tang, 2009). Meanwhile, these methods merged multi-modal features and ignored the similarity between news multi-modal data. The SAFE method (Zhou et al., 2020) pays attention to the similarity relationship between the image and the text. The similarity comparison module uses a pre-trained model image2sequence to complete the conversion of the image to the text (Vinyals, Toshev, Bengio, & Erhan, 2017). Compared with an emotional text, the goal of image2sequence is more inclined to the objective statement of image content. It lacks the emotional characteristics contained in images which is important for fake news detection.

Aiming at solving the problems of the above methods, this paper proposes MCNN. The MCNN is composed of five sub-networks: the text feature extraction module, the visual semantic feature extraction module, the visual tampering feature extraction module, the similarity measurement module, and the multimodal fusion module. Among them, the text feature extraction module and the visual semantic feature extraction module are responsible for extracting the vector representation of the semantic level of textual and visual features, and the visual tampering feature extraction module focuses on physical levels feature extraction such as malicious image tampering and recompression. The similarity measurement module can directly measure the similarity of news multimodal data. Our method can better capture the similarity of different modal data in multimodal news data, and the semantic features of texts and images are more suitable for fake news detection in complex scenes than existing methods.

The main contributions of this article are:

- We proposed a new neural network for fake news detection named MCNN, which using the similarity measurement module can measure the similarity of multimodal data of fake news (texts and images).
- To better capture the semantic features in the visual expression of fake news, we design a branch network for extracting visual semantic vectors, obtaining a better semantic expression of the picture.

- In the visual tampering feature extraction module, the ELA algorithm is introduced, and here, the Convolutional Neural Networks are employed to determine to judge the authenticity of fake news pictures at the physical level.
- Our method is a generic framework for fake news detection. These modules used for image and text feature extraction can be easily replaced by other models.

The structure of the article is organized as follows. Section 2 introduces related works of experts in correlative areas. Section 3 is the introduction and derivation of the method. The conclusions of experiments and prospects for future work are presented in Sections 4 and 5, respectively.

2. Related work

The detection of fake news has many related tasks, such as rumor detection (Cao et al., 2019; Imran, Ofli, Caragea, & Torralba, 2020; Ma et al., 2019) and spam detection (Kaghazgaran, Caverlee, & Squicciarini, 2018; Wang, Gong, & Fu, 2017; Wang, Liu, & Zhao, 2017). The main difficulty in fake news detection is finding the difference between the news based on features. We can obtain these features from forums, social environments and even from part of accompanying images, so we review the present work in the succeeding three areas: fake news detection based on traditional machine learning, fake news detection based on single modal data, and fake news detection based on multimodal data.

Methods based on hand-crafted features: This method mainly used hand-crafted features to detect fake news. They used feature engineering to extract the emotional polarity, user influence, geographic location, and similarity of dissemination structure in event-related information. Then they used these features to train decision trees, support vector machines, and other classifiers to classify events as fake news and real news (Castillo et al., 2011; Jin, Cao, Zhang, & Luo, 2016; Reis, Correia, Murai, Veloso, & Benevenuto, 2019; Wu, Yang, & Zhu, 2015). Castillo et al. (2011) used sentiment scores, including the quantity of URLs on Weibo, the quantity of days a user registered, and other characteristics to train the decision tree algorithm to detect rumors. Wu et al. (2015) adopted characteristics such as the geographical location involved in microblog, the client that issued microblog, and the emotional polarity of text symbols, etc., and then used a support vector machine classifier to detect rumors. Reis et al. (2019) evaluated 141 textual features and proposed a new set of features to detect fake news. However, designing effective hand-crafted features requires the knowledge of highly related area and specific events (Castillo et al., 2011; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Popat, Mukherjee, Strötgen, & Weikum, 2016). Meanwhile, this type of method relies on hand-crafted features and the robustness of the obtained feature vectors is not strong enough, since it lacks the knowledge of fake news detection. It is difficult to use hand-crafted features to detect fake news.

Single-modal methods based on deep learning: Today, many scholars have tried to use deep learning models to automatically construct deep features to detect fake news. Ma et al. (2016) aimed to find out the possibility of using deep neural networks to display tweets by capturing temporal-linguistic features. Chen, Li, Yin and Zhang (2018) added attentional mechanisms to recurrent neural networks (RNNs) to focus on different temporal-linguistic features with specific attention. The construction of deep learning models relies on plenty of labeled data, and the data acquisition of fake news has always been the main problem in the field of fake news detection. So the problem of data annotation has become the biggest bottleneck for rumor detection based on deep learning models. Some scholars tried to avoid data labeling and used the idea of unsupervised learning to detect online rumors. Chen, Zhang, Yeo, Lau and Lee (2018) proposed an unsupervised model which added multi-layer RNN to the front end of the autoencoder to detect rumors, it further improved the effectiveness of the model. Ángel González, Hurtado, and Pla (2020) proposed a contextualized pre-trained Twitter word embedding based model for irony detection via the transformer architecture. Although the unsupervised learning method avoids the problem of data labeling, the instability of the model brings greater limitations. Single-modal fake news detection based on deep learning improves accuracy compared to traditional methods but ignores news as a collection of multimedia data. The textual and visual information of fake news cannot be effectively used (Liu & Wu, 2018; Ma et al., 2016; Ma, Gao, & Wong, 2018).

Multimodality methods based on deep learning: In recent years, more scholars are focusing on deep learning methods based on multimodal data (Imran et al., 2020; Truong & Lauw, 2019). Zhao et al. (2019) proposed an image-text consistency driven multimodal approach to analyze the sentiment of social media. Liu, Zhang, and Gulla (2020) proposed a novel Attentive Recurrent Neural Network (Ante-RNN) with textual and visual fusion for the dynamic interpretable recommendation. Kumar, Srinivasan, Cheng, and Zomaya (2020) proffered a hybrid deep learning model for fine-grained sentiment prediction in real-time multimodal data. For the data of multiple modalities in news, current researchers use pre-trained deep Convolutional Neural Networks models such as VGG19 to extract the features of the image and merge the obtained visual performance with text information (Jin, Cao, Guo, Zhang and Luo, 2017; Khattar, Goud, Gupta, & Varma, 2019; Wang et al., 2018; Yang et al., 2018). Specifically, Jin, Cao, Guo et al. (2017) included social network multimodal content that used Deep Neural Networks to solve the problem of fake news detection. Wang et al. (2018) proposed an end-to-end event-based anti-neural network based on multimodal features to detect emerging fake news event. Khattar et al. (2019) proposed a new approach of learning shared representations for multimodal information for fake news detection. Nonetheless, these parts mainly aim at how to integrate different forms of information and ignore the effective modeling of visual content. These visual features used by them are so generic that it is incredibly difficult to indicate the internal features and the missing of fake news detection task-related information from the fake news image, which reduce the performance of visual content in the detection of fake news. At the same time, regarding text, the above models such as TextCNN or LSTM cannot uncover the connection between the text and context well, greatly reducing the ability of fake news detection in the text part (Jin, Cao, Guo et al., 2017; Zhou et al., 2020). Zhou et al. (2020) put forward an approach to detect fake news by comparing

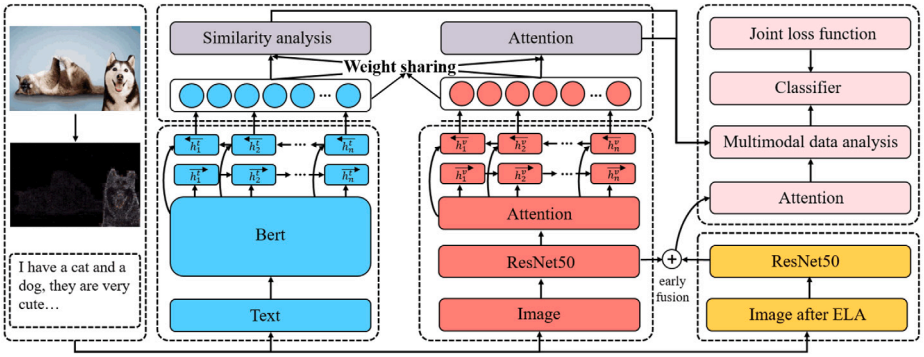

Fig. 2. The architecture of MCNN. The network in blue is the text feature extraction module. The network in red is the visual semantic feature extraction module. The network in orange is the visual tampering feature extraction module and the network in purple is the similarity measurement module. The network with the pink is the multimodal fusion module. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

text and image similarity, but the model uses a pre-trained picture description generation model and cannot calculate the similarity of multimodal data, which limits the scene used.

To address these limitations like designing hand-crafted features and general features of existing fake news detection methods, we propose a new network model (MCNN). Our model works jointly through five sub-networks, it captures well the similarity of different modal data in multimodal news data, semantic features of texts and images, and some physical features of the visual modalities. With our model, fake news detection in complex scenarios is more accurate than with existing models.

3. Methodlogy

3.1. Overview

To make the fake news detection more comprehensive, as shown in Fig. 2, we propose our fake news detection method MCNN. It consists of five modules, which are the text feature extraction module **(1)**, the visual semantic feature extraction module **(2)**, the visual tampering feature extraction module **(3)**, the similarity measurement module **(4)**, and the multimodal fusion module **(5)**.

Our research objective is to use the multimodal data (image and texts) of a news article to detect its authenticity. The problem is defined as follows: given an article of news $\boldsymbol{N}=\boldsymbol{(t,v)}$ containing both visual information ($\boldsymbol{v}$) and textual information ($\boldsymbol{t}$), we use $\boldsymbol{s}$ to denote the similarity between $\boldsymbol{t}$ and $\boldsymbol{v}$. Our goal is to use the visual information and textual information of a news $\boldsymbol{N}$ and the interconnection between them, i.e., to predict whether $\boldsymbol{N}$ is a piece of fake news ($\boldsymbol{y}=\boldsymbol{0}$) or a piece of true news ($\boldsymbol{y}=\boldsymbol{1}$), and their relationship to determine $(\boldsymbol{s,v,t}) \longrightarrow \mathbf{y} \in (\mathbf{0},\mathbf{1})$. The specific derivation of the MCNN is explained in the next sections.

3.2. The text feature extraction module

Most content-based fake news detection methods use the traditional word vector model, the traditional word vector model performs good in the modal analysis of short sentences and unambiguous sentences. However, most real sentences processed are not that simple. The problem of polysemy should be solved by considering the relationship between the words before and after. To solve this problem, we use the BERT pre-training model to extract text features (Devlin, Chang, Lee, & Toutanova, 2019), as shown in Eq. (1):

$$h_i^t = BERT(t_i) \tag{1}$$

where t_i represents the ith sentence of the input, and h_i^t represents the text features vector after BERT embedding.

Then to better capture global feature information and achieve better integration with image semantic information, we use BiGRU to extract the features extracted by BERT. BiGRU can further extract the temporal attributes of text features and turn text features into text feature sequences. As shown in Eq. (2):

$$f_i^t = BiGRU(h_i^t) \tag{2}$$

where h_i^t represents the text features vector after BERT embedding, and f_i^t represents the text features sequence extracted by BiGRU.

3.3. The visual semantic feature extraction module

Research (Qi et al., 2019) shows that fake news images tend to have stronger emotional factors than real news images, and the analysis of emotional factors is mainly reflected on the semantic and physical levels. In the stage of model designing, we take the output of the convolutional neural networks as the low-level features set of the image and use it to fuse it with the tampering detection module to realize the analysis of the image physical level. To obtain better semantic expression of the visual part, we first use the ResNet50 pre-training model to encode the input image. Before the classification layer of the pre-trained ResNet50 model, we used a 1024-dimensional fully connected layer to encode the image features. The image representation is a 1024-dimensional vector. The process is shown in Eq. (3):

$$h^v = ResNet50(v) \tag{3}$$

where v represents the original image of input, and h^v represents the visual semantic features extracted by ResNet50.

Afterwards we feed the semantic features of the image to the attention mechanism to highlight the regions of the image that have strong emotional expression (Vaswani et al., 2017). Therefore, when a visual modal representation is obtained, each feature will be assigned a weight to indicate its "importance" in the modal representation. As shown in Eqs. (4)–(6):

$$u^v = U^T \tanh(W^v h^v + b^v) \tag{4}$$

$$\alpha^v = \frac{\exp(u^v)}{\sum_i \exp(u^v)} \tag{5}$$

$$s^v = \sum_i \alpha_i h^v \tag{6}$$

where W^v represents a weight matrix, b^v is a bias term, U^T represents a transposed weight vector, and u is a scoring function that evaluates the importance of every single eigenvector. After that, the softmax function is used to obtain the normalized weight of the ith eigenvector u^v, and the hierarchy of the input image is calculated as the weighted sum of different eigenvectors. During training, vectors are randomly initialized and learned together. So far, we have got an advanced semantic representation of the input image.

To obtain a better semantic expression of the image (Lang, 1979), we used BiGRU to form the image to sequence module. The image to sequence module is often used for image description generation to align image features with text features. We introduce the image to sequence module here and send the feature expression of the picture to BiGRU to obtain the semantic vector of the visual modality. This step is equivalent to the commonly used embedding layer in text analysis, which transforms the semantic feature analysis of the image to the level of the semantic sequence.

The representation of the proposed method is shown in Eq. (7). Compared with directly using the features of the image, this method is more helpful to express the semantic information of the image.

$$f^v = BiGRU(s^v) \tag{7}$$

where s^v represents the advanced semantic representation extracted by ResNet50, and f^v represents the visual semantic sequence extracted by BiGRU.

3.4. The visual tampering feature extraction module

Compared with the real news image, the fake news image is often maliciously spliced or the number of times of recompression is increased due to the number of times of propagation. We found that the Error Level Analysis (ELA) algorithm can highlight the malicious stitching and recompression characteristics of fake images better than directly transforming the image into the frequency domain through the experiment. As shown in Fig. 3. Meanwhile, we use svd to compress the image, with a compression rate of 0.7 each time, and then use the ELA algorithm to process the compressed image. It can be seen that with the different compression times, the image processed by the ELA algorithm is gradually changing. The results are presented in Fig. 4. Which shows that ELA algorithm can highlight the malicious stitching and recompression characteristics of fake images.

In the visual tampering feature extraction module, we use the ELA transformation of the image firstly, and then apply the ResNet50 model to extract the image tampering features, as shown in Eqs. (8)–(9):

$$v^{ela} = ELA(v) \tag{8}$$

$$h^{ela} = ResNet50(v^{ela}) \tag{9}$$

where v represents the original image of input, v^{ela} represents the original image processed with ELA, and h^{ela} represents the tampered features extracted by ResNet50.

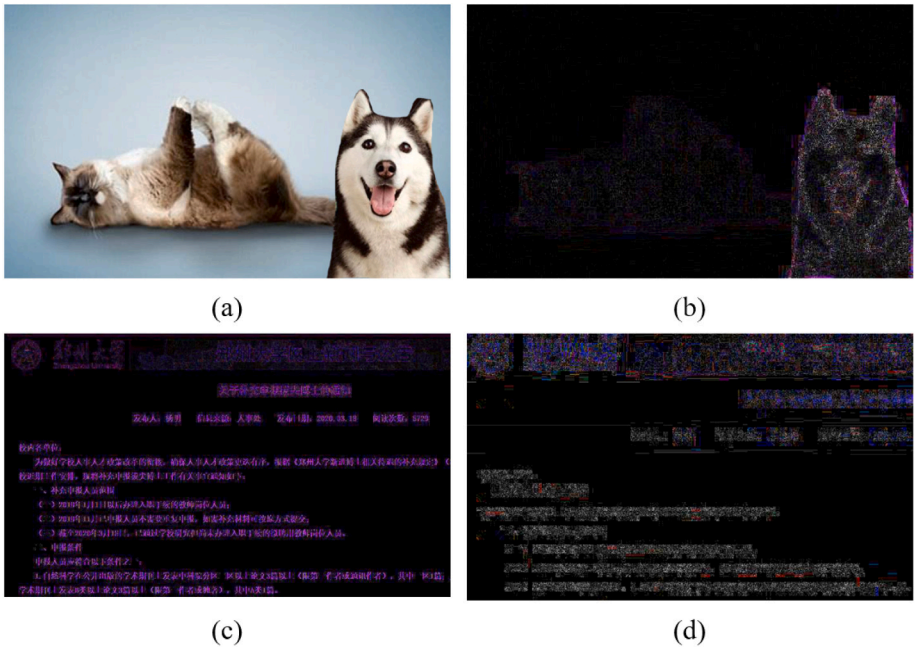

Fig. 3. Image processed by ELA. Among them, (a) is a tampered image, (b) is (a) after ELA processing, and the Husky is highlighted as a tampered region. (c) is the image that has not been recompressed and has undergone ELA processing. (d) is (c) after ELA processing after re-compression transformation, we can see that the re-compressed image and the original image show different characteristics after ELA transformation.

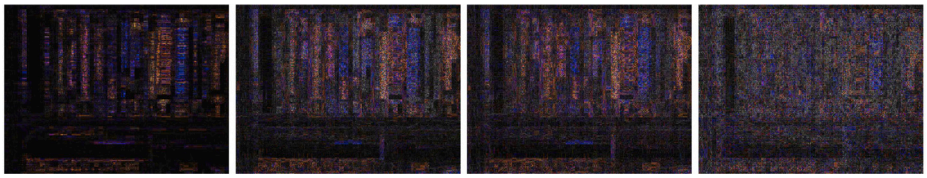

Fig. 4. The image processed by the ELA algorithm with different re-compress times.

3.5. The similarity measurement module

In fake news detection, besides employing the feature fusion of the fake news directly, the falsity of news articles can also be detected by evaluating the correlation between the text information and its visual information. We design a similarity measurement module to measure the similarity of fake news texts and images directly. It is described next.

In the previous module, we have obtained the vector representation of texts and images through the visual semantic feature extraction sub-network and the text semantic feature sub-network. To certify that the two sub-networks learn the common representation space of image and text patterns, we apply a fully connected layer to the last layer of each sub-network and force the two sub-networks to share weights of the last layer (Zhen, Hu, Wang, & Peng, 2019). We obtain the semantic representation s^t, s^v of the image and text after sharing features. This is to visually produce as similar a representation as possible for the same category of image and text samples. Afterwards we apply cosine similarity to measure the similarity between image and text. As shown in Eq. (10):

$$s = \frac{s^t \cdot s^v}{\|s^t\| \times \|s^v\|} \tag{10}$$

where s^t, s^v represents image semantic sequence and text feature sequence respectively.

Here the s takes a range of [−1,1], the larger value represents the higher similarity between the text and image. To map the similarity between 0 and 1, we choose the sigmoid activation function here to map the similarity between [0,1], as shown in Eq. (11):

$$p^s = sigmoid(s) \tag{11}$$

where sigmoid is the activation function used to map the similarity between 0 and 1.

We suppose that news articles formed by a mismatch between textual and visual information can be falsified much easier than news articles formed by matching image and text, analyzed from a pure similarity perspective. Next we can establish a loss function

based on cross-entropy as follows:

$$\mathcal{L}_s\left(\theta_t,\theta_v\right)=-\mathbb{E}_{(a,y)\sim(A,Y)}\left(y\log\left(1-p^s\right)+(1-y)\log p^s\right) \tag{12}$$

$$\left(\hat{\theta}_t,\hat{\theta}_v\right)=\arg\min_{\theta_t,\theta_v}\mathcal{L}_s\left(\theta_t,\theta_v\right) \tag{13}$$

3.6. The multimodal fusion module

Through the four previous subnetworks, we obtain the feature representation h^v, h^{ela} at the physical level of the image, and the feature representation s^c=[s^v, s^t] at the semantic level of the image and text. In the multimodal fusion module, we use the attention mechanism to assign weights to the physical level features and the semantic level features of the image and text which can be expressed as f^c = [s^c, f^p, h^v, h^{ela}] to highlight more valuable features, as shown in Eqs. (14)–(16):

$$u^c=U^T\tanh\left(W^c f_i^c+b^c\right) \tag{14}$$

$$\alpha_i=\frac{\exp\left(u^c\right)}{\sum_i\exp\left(u^c\right)} \tag{15}$$

$$s_i^e=\sum_i\alpha_i f_i^c \tag{16}$$

where f^p represents the physical features of the image, W^c represents a weight matrix, b^c is a bias term, U^T represents a transposable weight vector, and u is a scoring function that weights the importance of each eigenvector. Meanwhile, through this step, we get the fused features of image and text which are assigned to attention weights.

Our goal is to map the textual and visual features derived from news to their tags, thereby predicting the probability that they are fake news. We apply the softmax function to implement the correspondence between features and labels, as shown in Eq. (17):

$$p^c=softmax\left(\mathbf{W}_p\cdot s^e+\mathbf{b}_p\right) \tag{17}$$

we also define a loss function based on cross entropy: $\mathcal{L}_p(\theta_t,\theta_v,\theta_p)=-\mathbb{E}_{(a,y)\sim(A,Y)}(y\cdot\log p^c+(1-y)\cdot\log p^c)$. Our goal is to correctly identify fake news through news multimodal feature fusion and similarity of news multimodal data, and to involve both cases, we specify the final loss function as shown in Eq. (18), α and β means the loss function weights of the two branch networks and $\alpha+\beta=1$.

$$\mathcal{L}\left(\theta_t,\theta_v,\theta_p\right)=\alpha\mathcal{L}_p\left(\theta_t,\theta_v,\theta_p\right)+\beta\mathcal{L}_s\left(\theta_t,\theta_v\right) \tag{18}$$

where parameters can be jointly learned by:

$$\left(\hat{\theta}_t,\hat{\theta}_v,\hat{\theta}_p\right)=\arg\min_{\theta_t,\theta_v,\theta_p}\mathcal{L}\left(\theta_t,\theta_v,\theta_p\right) \tag{19}$$

4. Experiments

In this section, we show experiments with datasets from four real scenarios to measure the validity of the constructed MCNN. We also answer the evaluation questions listed below:

- **EQ1:** Can MCNN outperform other methods in the task of fake news detection?
- **EQ2:** Are the five modules in the model effective in detecting fake news?

In this section, we initially described datasets proposed by us and show some of the baseline methods used to detect fake news. And then we compared MCNN to those baselines and an ablation study that was established to answer the EQ1 and EQ2 respectively. Finally, we examined some typical cases to show the importance of multi-domain detection of fake news images.

4.1. Datasets

To assess the effect of the method fairly, the experiments were established on four real social media datasets. The datasets are described in detail as follows:

- **D1:** Yang et al. (2018) proposed the dataset (D1), which includes 20,015 news, i.e. 11,941 fake ones and 8,074 real ones. For fake News in D1, it contains text and metadata scraped from more than 240 websites by the Megan Risdal on Kaggle. The 315 real items were obtained from the New York Times, Washington Post, etc. In the experiments of this article, we only used the news that contains visual information.
- **D2:** MCG-FNeWS (D2) is the largest Chinese fake news dataset currently publicly available. It was proposed by the Institute of Computer Technology, Chinese Academy of Sciences. This dataset covers 19186 non-fake news and 19258 fake news published on Weibo (https://weibo.com) from May 2012 to November 2018. It was also used for the Zhiyuan Fake News Recognition Competition (Cao et al., 2019).

Table 1
The statistics of the datasets.

Dataset	D1	D2	D3	D4
# of fake news	11 941	19 258	7898	320
# of real news	8074	19 186	6026	528
# of images	6529	34 096	514	683

- **D3:** The Twitter dataset(D3) was one of the components of MediaEval (Maigrot, Claveau, Kijak, & Sicre, 2016), which was applied to validate the usage task of Multimedia as well as aims to detect the fake multimedia content on social media. The dataset is composed of tweets (short messages posted on Twitter) and each tweet has either text, image/video or social context information.
- **D4:** (Shu, Sliva, Wang, Tang, & Liu, 2017) PolitiFact (D4) (**politifact.com**), it is a well-known non-profit political statement and website that reports fact-checking in the United States. The news articles in the PolitiFact dataset were published from May 2002 to July 2018. The tags of news articles in this dataset (fake or true) are provided by domain experts to ensure the quality of news tags.

The statistical data for the datasets as Table 1 shown:

4.2. Baselines

The disadvantages of existing multimodal methods for the fake news analysis on multimodal data are that they only directly stitch the image and text features of the news without considering the similarity of multimodal data, which leads to the mismatch of the image and text in the fake news cannot be recognized properly. Moreover, these methods only incorporate the semantic features of the image, which cannot effectively identify the malicious tampering image in the fake news.

To solve the problem about the limitations of existing works, our method works jointly through five sub-networks. It can capture the similarity of different modal data in multimodal news data, semantic level features of texts and images, and some physical level features of the visual modalities. The main advantages of our proposed method are as follows:

- The similarity measurement module that can measure the similarity of the news with multimodal data directly, which can detect the news with mismatch images.
- The visual tampering feature extraction module can detect the malicious tampered images of fake news effectively.

To evaluate the performance of the existing methods, the following baselines were applied. *S* means single-modal and *M* means multimodal.

1. **(S) LIWC:** LIWC (Pennebaker, Boyd, Jordan, & Blackburn, 2015) is a widely-accepted psycho-linguistics lexicon. LIWC can count the words in the text falling into one. These numbers act as hand-crafted features used by random forest to predict fake news.
2. **(S) Visual:** Pre-trained VGG-19 and a fully connected layer are utilized to get the visual feature RV. Afterward RV is sent to make the prediction into a 32-dimensional fully connected layer (Simonyan & Zisserman, 2015). The VGG-19 network we used is a pre-trained model with imagenet weights.
3. **(M) VQA:** The Visual Question Answering (VQA) (Antol et al., 2015) model is used to answer the questions based on the figures given before. The initial VQA model is designed for multi-class classification tasks. We are going to explore binary classification in this section. We used the feature extraction part of the VQA model, meanwhile, one-layer LSTM with 32-dimension is used.
4. **(M) att-RNN:** The att-RNN (Jin, Cao, Guo et al., 2017) is a kind of framework based on multimodal fake news detection, the text, visual and social context information are fused in it, which applies LSTM and VGG19 to extract the textual and visual features respectively in the stage of modal structure, and also fuses the features we obtained through the attention mechanism. This paper, the att-RNN is compared by us to test the accuracy of fake news detection.
5. **(M) EANN:** The EANN (Wang et al., 2018) contains three major parts: the multimodal feature extractor, the fake news detector and the event discriminator. The multimodal feature extractor acquires textual and visual features from the posts. To detect fake news, we only used the multimodal feature extractor and the fake news detector. Meanwhile, the configure of EANN is set as the official implementation.[1]
6. **(M) MVAE:** The MVAE (Khattar et al., 2019) uses a bimodal variational autoencoder coupled with a binary classifier to finish the detection of fake news. It uses LSTM with 32-dimension in textual encoder, and the visual encoder uses two fully connected layers of size 1024 and 32. The fake news detector has a 64-dimension fully connected layer and a 32 fully connected layer. We use the official implementation of MVAE.[2]

[1] https://github.com/yaqingwang/EANN-KDD18.
[2] https://github.com/dhruvkhattar/MVAE.

Table 2
The results of different methods on four datasets. The highest score is highlighted in bold.

		LIWC	Visual	VQA	att-RNN	EANN	MVAE	SAFE	BERT+MVNN	Ours
D1	Acc.	0.796	0.758	0.778	0.899	0.855	0.908	0.922	0.956	**0.963**
	Pre.	0.807	0.761	0.773	0.904	0.863	0.915	0.937	0.964	**0.972**
	Rec.	0.789	0.752	0.784	0.883	0.843	0.909	0.935	0.959	**0.964**
	F1	0.798	0.756	0.778	0.894	0.852	0.912	0.936	0.961	**0.968**
D2	Acc.	0.774	0.743	0.763	0.783	0.823	0.876	0.924	0.940	**0.947**
	Pre.	0.780	0.754	0.772	0.789	0.836	0.883	0.939	0.941	**0.952**
	Rec.	0.772	0.743	0.773	0.792	0.824	0.873	0.930	0.943	**0.942**
	F1	0.776	0.748	0.773	0.790	0.830	0.878	0.929	0.941	**0.946**
D3	Acc.	0.684	0.596	0.631	0.664	0.715	0.743	0.762	0.769	**0.784**
	Pre.	0.703	0.695	0.765	0.749	0.822	0.832	0.831	0.828	**0.850**
	Rec.	0.692	0.518	0.509	0.615	0.638	0.784	0.822	0.814	**0.814**
	F1	0.697	0.593	0.611	0.676	0.719	0.807	0.823	0.821	**0.831**
D4	Acc.	0.822	0.649	0.705	0.769	0.759	0.812	0.874	0.880	**0.884**
	Pre.	0.785	0.668	0.723	0.735	0.764	0.803	0.889	0.948	**0.973**
	Rec.	0.846	0.787	0.764	**0.942**	0.806	0.835	0.903	0.893	0.867
	F1	0.815	0.720	0.743	0.826	0.784	0.819	0.896	0.919	**0.917**

7. **(M) SAFE:** Zhou et al. (2020) proposed a method to detect fake news through the similarity between texts and images. It uses a pre-trained image to text model to transform the image into text, and then measure the similarity. We reduced the official implementation of SAFE.[3]
8. **(M) BERT+MVNN:** We used BERT to classify the news text, and MVNN is used to classify the corresponding news image. After each of the modalities has made a result, we used late fusion to get the final result.

Next, we will introduce the experimental design in detail.

4.3. Implementation details

In this section, we present the implementation details of MCNN. In the text feature extraction module, we use BERT to encode the text. Firstly, we set the length of the input text to 16. The input text is formatted according to the tokenize of BERT. Then a BiGRU layer is used for align with image features and better express the semantic features of the text. The number of BiGRU hidden layers to 256. In the visual semantic feature extraction module, the input size of the image is 224 × 224, and we use the output of the ResNet50 pre-trained on ImageNet set. The weights of the ResNet50 are frozen. After passing ResNet50, we followed by a fully connected layer with a dimension of 1024. In order to establish the correlation of internal modes and align with the text features, we use BiGRU with dimension 256 to accept the feature expression of the image. The layers of BiGRU is a hyperparameter, and we find that 1 is the most suitable and not easy to over fit. Each BiGRU layer is followed by a dropout layer with a rate of 0.4. The visual tampering feature extraction module uses the same ResNet50 network as the visual semantic feature extraction module. In the similarity measurement module, we set up a fully connected layer with shared parameters of dimension 256 for image and text. In the joint loss function, we set α and β to 0.7 and 0.3 respectively.

The datasets are divided into training set, validation set and test set according to the ratio of 7:1:2. In the ablation analysis, we retrained each branch network after remove parts of them. We use the batch size of 32 and the model is trained for 100 epochs with a learning rate 10-4. Also the early stopping is used to avoid overfitting. ReLU is used as the non-linear activation function. In order to get optimal parameters for our model, we use Adam as the optimizer.

4.4. Performance comparison

In this section, various experiments are constructed to compare the performance between the present baselines and the MCNN. We use the Accuracy, F1 Acore, Precall and Recall of the fake-news class as evaluation metrics **(EQ1)**. The formula of evaluation index is shown in Eqs. (20)–(23):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

$$Precall = \frac{TP}{TP + FP} \tag{21}$$

$$Recall = \frac{TP}{TP + FN} \tag{22}$$

$$F1 = \frac{2 \times Precall \times Recall}{Precall + Recall} \tag{23}$$

[3] https://github.com/Jindi0/SAFE.

Table 3
Architecture ablation analysis of MCNN.

	Part1	Part2	Part3	Part4	Acc.	Pre.	Rec.	F1.
D1	✓				0.948	0.957	0.951	0.954
	✓	✓			0.950	0.962	0.950	0.957
	✓	✓	✓		0.958	0.968	0.960	0.963
	✓	✓	✓	✓	0.963	0.972	0.964	0.968
D2	✓				0.926	0.931	0.920	0.925
	✓	✓			0.936	0.939	0.932	0.936
	✓	✓	✓		0.941	0.940	0.942	0.940
	✓	✓	✓	✓	0.947	0.952	0.942	0.946
D3	✓				0.742	0.811	0.775	0.793
	✓	✓			0.753	0.820	0.782	0.801
	✓	✓	✓		0.771	0.835	0.806	0.820
	✓	✓	✓	✓	0.784	0.850	0.814	0.831
D4	✓				0.857	0.959	0.845	0.899
	✓	✓			0.866	0.962	0.855	0.904
	✓	✓	✓		0.875	0.969	0.865	0.910
	✓	✓	✓	✓	0.884	0.973	0.867	0.917

where TP is the number of positive samples predicted by the model to be positive, TN is the number of negative samples predicted by the model to be negative, FP is the number of negative samples predicted by the model to be positive, and FN is the number of positive samples predicted by the model to be negative.

The comparison results between the proposed method and the baseline methods are shown in Table 2. From Table 2, the proposed MCNN presents better than the baseline in accuracy, precision, recall and F1 score. From the experimental results, we can see that the fake news recognition method based on a single modality is much weaker than the fake news recognition method based on multimodal data. Although visual modalities are effective for fake news detection, the performance based on a single visual modal is still worse than the multimodal method. This confirms that the method of integrating multimodal features is suitable for fake news detection. In multimodal models, the performance of att-RNN is better than VQA, which also indicates that the application of attention mechanism can perfect the model's performance. Compared with other methods, BERT+MVNN has a higher accuracy rate, but it is weaker than MCNN. The reason is that although MVNN pays attention to the fusion of multiple visual features, the method based on late fusion cannot consider the difference between the visual modality and the textual.

The proposed method in this paper improves the detection accuracy on four datasets compared with the best performing methods. Improved results on four datasets are D1 = D2 > D3 > D4. Meanwhile, the size relationship of the four datasets are D2 > D1 > D3 > D4. It can be seen that our method perform better on larger datasets. The reason for this result is large data sets have stronger data diversity and can substantiate the training of the method proposed in this article.

4.5. Architecture ablation analysis

In this section, our goal from the perspective of quantitative and qualitative assessment areas and other network components is improving MCNN **(EQ2)** and the effectiveness of the proposed method. We conduct an ablation analysis, starting with the most basic configuration and incrementing the components that build the complete architecture. The results are listed in Table 3.

We start with the text feature extraction module, and we go from relying only on text features to adding new subnetworks in turn. As shown in the first line of Table 3, the average accuracy achieved is 94.8%, 92.6%, 74.2%, 85.7% only depending on the characteristics of the text (Part1). Then we add the visual semantic feature extraction module (Part2), and we can see that the effect is improved by 0.2%, 1.1%, 1.4% and 1.1%, respectively compared with the single module. Then we add the similarity measurement module (Part3), and we can see that the average recognition accuracy is achieved 95.8%, 94.1%, 77.1%, 87.5%. Compared with part1+part2, the effect is improved by 0.8%, 0.5%, 1.8% and 0.9%. This also proves that the subnetwork we proposed directly used for the similarity measurement of multimodal data is effective. Finally, we add the multimodal fusion module (Part4), to effectively merge the physical layer features of the image and used the attention mechanism to allocate the weights of the physical level features of the image and text as well as the image. Meanwhile, to combine the similarity measurement module. Here, the experimental results have reached the best performance 96.3%, 94.7%, 78.4%, 88.4%. The results raise 0.5%, 0.6%, 1.3% and 0.9% respectively.

To better illustrate the ablation experiment better, we plotted the results. As proven in Fig. 5, it can be clearly seen from the figure that the accuracy of the correspondence increases with each module in the model.

4.6. Case study

To show the importance of applying the multimodal data in fake news detection more clearly, we conduct an individual case analysis on the prediction results from the MCNN. Thus, the function of the MCNN can be more objectively expressed.

We remove the visual tampering feature extraction module from the complete MCNN. Then, we compare the results before and after removal, and find that the news images shown in Fig. 6 are not accurately identified after removing the visual tampering

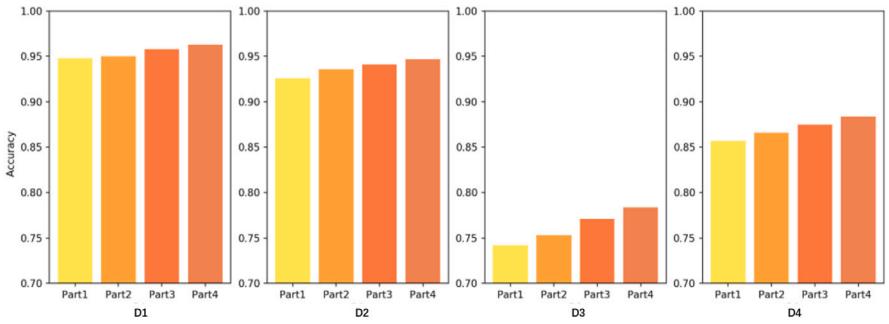

Fig. 5. The result of ablation analysis. We can visually see that the detection accuracy increases as the subnetwork increases, which means our subnetwork can cooperate work to detect fake news.

Fig. 6. Example of news detected by the visual tampering feature extraction module.

feature extraction module. By analyzing its ELA images, we can intuitively find that these images show obvious tampering features. Therefore, after removing the visual tampering feature extraction module, the MCNN cannot capture visual tampering feature information, resulting in that the fake news cannot be correctly detected.

5. Conclusion

This paper studies the problem of multimodal fake news detection. Most existing methods focus on directly splicing image features and text features and cannot fully identify fake news. We innovatively proposed MCNN. Our method consists of five parts, namely the text feature extraction module, the visual semantic feature extraction module, the visual tampering feature extraction module, the similarity measurement module, and the multimodal fusion module. Through the joint work of these five modules, our method can better capture the similarity of different modal data in multimodal news data, the semantic level features of texts and images, and some physical level features of visual modal. Compared with existing methods, it is more suitable for fake news detection in complex scenes. At the same time, the experiments are conducted on four widely used datasets, and the results of the experiments indicated that our method is ahead of the existing baseline method in detecting fake news based on multimodal data. This also proves the effectiveness of our method.

In the future, we will continue to optimize the method at the level of feature fusion to make multimodal features from different sources more fit. We will also apply this method to other complex problem scenarios based on multimodal data.

Acknowledgment

We would like to thank the anonymous reviewers for their constructive comments. This paper is supported by the plan for Young Backbone Teachers in Henan Province (No. 22020GGJS014).

References

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236. http://dx.doi.org/10.1257/jep.31.2.211.

Ángel González, J., Hurtado, L.-F., & Pla, F. (2020). Transformer based contextualization of pre-trained word embeddings for irony detection in Twitter. *Information Processing & Management*, *57*(4), Article 102262. http://dx.doi.org/10.1016/j.ipm.2020.102262.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., et al. (2015). VQA: Visual question answering. In *ICCV'15*, *International conference on computer vision* (pp. 2425–2433). Santiago, Chile: http://dx.doi.org/10.1109/ICCV.2015.279.

Cao, J., Sheng, Q., Qi, P., Zhong, L., Wang, Y., & Zhang, X. (2019). False news detection on social media. arXiv:1908.10818.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. In *WWW '11*, *Proceedings of the 20th international conference on world wide web* (pp. 675–684). http://dx.doi.org/10.1145/1963405.1963500.

Chen, T., Li, X., Yin, H., & Zhang, J. (2018). Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In *PAKDD'18*, *Trends and applications in knowledge discovery and data mining* (pp. 40–52). Cham: http://dx.doi.org/10.1007/978-3-030-04503-6_4.

Chen, W., Zhang, Y., Yeo, C. K., Lau, C. T., & Lee, B. S. (2018). Unsupervised rumor detection based on users' behaviors using neural networks. *Pattern Recognition Letters*, *105*, 226–233. http://dx.doi.org/10.1016/j.patrec.2017.10.014.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *ACL'19*, *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies* (pp. 4171–4186). Minneapolis, Minnesota: http://dx.doi.org/10.18653/v1/N19-1423.

Imran, M., Ofli, F., Caragea, D., & Torralba, A. (2020). Using AI and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions. *Information Processing & Management*, *57*(5), Article 102261. http://dx.doi.org/10.1016/j.ipm.2020.102261.

Jin, Z., Cao, J., Guo, H., Zhang, Y., & Luo, J. (2017). *MM '17*, *Multimodal fusion with recurrent neural networks for rumor detection on microblogs* (pp. 795–816). Mountain View, California, USA: http://dx.doi.org/10.1145/3123266.3123454.

Jin, Z., Cao, J., Zhang, Y., & Luo, J. (2016). News verification by exploiting conflicting social viewpoints in microblogs. In *AAAI'16*, *Proceedings of the thirtieth AAAI conference on artificial intelligence* (pp. 2972–2978). Phoenix, Arizona: http://dx.doi.org/10.5555/3016100.3016318.

Jin, Z., Cao, J., Zhang, Y., Zhou, J., & Tian, Q. (2017). Novel visual and statistical image features for microblogs news verification. *IEEE Transactions on Multimedia*, *19*(3), 598–608. http://dx.doi.org/10.1109/TMM.2016.2617078.

Kaghazgaran, P., Caverlee, J., & Squicciarini, A. (2018). Combating crowdsourced review manipulators: A neighborhood-based approach. In *WSDM'18*, *Proceedings of the eleventh ACM international conference on web search and data mining* (pp. 306–314). Marina Del Rey, CA, USA: http://dx.doi.org/10.1145/3159652.3159726.

Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019). MVAE: Multimodal variational autoencoder for fake news detection. In *WWW '19*, *The world wide web conference* (pp. 2915–2921). San Francisco, CA, USA: http://dx.doi.org/10.1145/3308558.3313552.

Kumar, A., Srinivasan, K., Cheng, W.-H., & Zomaya, A. Y. (2020). Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. *Information Processing & Management*, *57*(1), Article 102141. http://dx.doi.org/10.1016/j.ipm.2019.102141.

Lang, P. J. (1979). A bio-informational theory of emotional imagery. *Psychophysiology*, *16*(6), 495–512. http://dx.doi.org/10.1111/j.1469-8986.1979.tb01511.x.

Lin, Z., He, J., Tang, X., & Tang, C.-K. (2009). Fast, automatic and fine-grained tampered JPEG image detection via dct coefficient analysis. *Pattern Recognition*, *42*(11), 2492–2501. http://dx.doi.org/10.1016/j.patcog.2009.03.019.

Liu, Y., & Wu, Y. B. (2018). Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the thirty-second AAAI conference on artificial intelligence* (pp. 354–361).

Liu, P., Zhang, L., & Gulla, J. A. (2020). Dynamic attention-based explainable recommendation with textual and visual fusion. *Information Processing & Management*, *57*(6), Article 102099. http://dx.doi.org/10.1016/j.ipm.2019.102099.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., et al. (2016). Detecting rumors from microblogs with recurrent neural networks. In *IJCAI'16*, *Proceedings of the twenty-fifth international joint conference on artificial intelligence* (pp. 3818–3824). New York, NY, USA: http://dx.doi.org/10.5555/3061053.3061153.

Ma, J., Gao, W., & Wong, K.-F. (2018). Rumor detection on Twitter with tree-structured recursive neural networks. In *ACL'18*, *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 1980–1989). Melbourne, Australia: http://dx.doi.org/10.18653/v1/P18-1184.

Ma, J., Gao, W., & Wong, K.-F. (2019). Detect rumors on Twitter by promoting information campaigns with generative adversarial learning. In *WWW '19*, *The world wide web conference* (pp. 3049–3055). New York, NY, USA: http://dx.doi.org/10.1145/3308558.3313741.

Maigrot, C., Claveau, V., Kijak, E., & Sicre, R. (2016). Mediaeval 2016: A multimodal system for the verifying multimedia use task. In *MediaEval*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *NIPS'13*, *Proceedings of the 26th international conference on neural information processing systems - Volume 2* (pp. 3111–3119). Lake Tahoe, Nevada: http://dx.doi.org/10.5555/2999792.2999959.

Pennebaker, J. W., Boyd, R. L., Jordan, K. N., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015.

Popat, K., Mukherjee, S., Strötgen, J., & Weikum, G. (2016). Credibility assessment of textual claims on the web. In *CIKM '16*, *Proceedings of the 25th ACM international on conference on information and knowledge management* (pp. 2173–2178). Indianapolis, Indiana, USA: http://dx.doi.org/10.1145/2983323.2983661.

Qi, P., Cao, J., Yang, T., Guo, J., & Li, J. (2019). Exploiting multi-domain visual information for fake news detection. In *ICDM'19*, *2019 IEEE international conference on data mining* (pp. 518–527). Macau, China: http://dx.doi.org/10.1109/ICDM.2019.00062.

Qian, F., Gong, C., Sharma, K., & Liu, Y. (2018). Neural user response generator: Fake news detection with collective user intelligence. In *IJCAI'18*, *Proceedings of the 27th international joint conference on artificial intelligence* (pp. 3834–3840). AAAI Press, http://dx.doi.org/10.5555/3304222.3304302.

Reis, J. C. S., Correia, A., Murai, F., Veloso, A., & Benevenuto, F. (2019). Supervised learning for fake news detection. *IEEE Intelligent Systems*, *34*(2), 76–81. http://dx.doi.org/10.1109/MIS.2019.2899143.

Rubin, V., Conroy, N., Chen, Y., & Cornwell, S. (2016). Fake news or truth? Using satirical cues to detect potentially misleading news. In *ACL'16*, *Proceedings of the second workshop on computational approaches to deception detection* (pp. 7–17). San Diego, California: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/W16-0802.

Ruchansky, N., Seo, S., & Liu, Y. (2017). Csi: A hybrid deep model for fake news detection. In *CIKM '17*, *Proceedings of the 2017 ACM on conference on information and knowledge management* (pp. 797–806). http://dx.doi.org/10.1145/3132847.3132877.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *SIGKDD Exploration Newsletters*, *19*(1), 22–36. http://dx.doi.org/10.1145/3137597.3137600.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

Sunstein, C. R. (2009). On rumors: How falsehoods spread, why we believe them, and what can be done.

Truong, Q.-T., & Lauw, H. (2019). Multimodal review generation for recommender systems. In *WWW '19*, *The world wide web conference* (pp. 1864–1874). San Francisco, CA, USA: http://dx.doi.org/10.1145/3308558.3313463.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *NIPS'17*, *Proceedings of the 31st international conference on neural information processing systems* (pp. 6000–6010). Long Beach, California, USA: http://dx.doi.org/10.5555/3295222.3295349.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2017). Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(04), 652–663. http://dx.doi.org/10.1109/TPAMI.2016.2587640.

Wang, B., Gong, N. Z., & Fu, H. (2017). GANG: Detecting fraudulent users in online social networks via guilt-by-association on directed graphs. In *ICDE'17*, *2017 IEEE international conference on data mining* (pp. 465–474). New Orleans, LA, USA: http://dx.doi.org/10.1109/ICDM.2017.56.

Wang, X., Liu, K., & Zhao, J. (2017). Handling cold-start problem in review spam detection by jointly embedding texts and behaviors. In *ACL'17*, *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 366–376). Vancouver, Canada: http://dx.doi.org/10.18653/v1/P17-1034.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., et al. (2018). EANN: Event adversarial neural networks for multi-modal fake news detection. In *KDD '18, Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery amp; data mining* (pp. 849–857). London, United Kingdom: http://dx.doi.org/10.1145/3219819.3219903.

Wu, K., Yang, S., & Zhu, K. Q. (2015). False rumors detection on sina weibo by propagation structures. In *ICDE'15, 2015 IEEE 31st international conference on data engineering* (pp. 651–662). Seoul, South Korea: http://dx.doi.org/10.1109/ICDE.2015.7113322.

Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z., & Yu, P. S. (2018). TI-CNN: convolutional neural networks for fake news detection. CoRR abs/1806.00749 arXiv:1806.00749 URL http://arxiv.org/abs/1806.00749.

Zhao, Z., Zhu, H., Xue, Z., Liu, Z., Tian, J., Chua, M. C. H., et al. (2019). An image-text consistency driven multimodal sentiment analysis approach for social media. *Information Processing & Management, 56*(6), Article 102097. http://dx.doi.org/10.1016/j.ipm.2019.102097.

Zhen, L., Hu, P., Wang, X., & Peng, D. (2019). Deep supervised cross-modal retrieval. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 10386–10395). http://dx.doi.org/10.1109/CVPR.2019.01064.

Zhou, X., Wu, J., & Zafarani, R. (2020). SAFE: Similarity-aware multi-modal fake news detection. In *PAKDD '20, Pacific-asia conference on knowledge discovery and data mining* (pp. 354–367). Singapore, SG: Springer, http://dx.doi.org/10.1007/978-3-030-47436-2_27.