Multimodal integration for fake news detection on social media platforms

Facheng Yan*, *Mingshu* Zhang, and *Bin* Wei

Engineering University of PAP, Xian, China

Keywords: Fake news detection deep learning, Cybersecurity.

Abstract. The widespread dissemination of fake news on social media platforms can cause serious social impact, making the detection of fake news on social media platforms an urgent problem to be solved. Up to now, scholars have proposed various methods ranging from traditional manual feature extraction to deep learning algorithms for detecting fake news. However, these methods still have some limitations and two difficult problems: (1) How to learn informative news feature representations without losing information as much as possible? (2) How to effectively fuse multi-modal information to obtain high-order complementary information about news and enhance fake news detection? To overcome these two difficulties, this article proposes a multi-modal fusion fake news detection model. Firstly, the model uses BERT and VGG-19 to obtain the text and image feature representations of news content, respectively, and then further fuses multi-modal information through a multi-modal attention mechanism module to obtain high-order complementary information between different modalities, thereby obtaining informative news feature representations for fake news detection. Experimental results on two real-world public datasets demonstrate the effectiveness of our model compared to mainstream detection methods.

1 Introduction

With the rapid development of social networks, the way people obtain information is also changing. Twitter, Facebook, Sina Weibo and other emerging social media platforms have become the main channels for the public to obtain news [1]. Due to the strong openness of emerging media platforms, users can post or repost news articles at will. Therefore, tens of thousands of news articles are widely disseminated on social media platforms every day. However, due to the randomness of news posting and the lack of verification and inspection of these news articles by various institutions, all kinds of fake news emerge endlessly on social media, which will bring tremendous political, economic and social public opinion influence.

* Corresponding author: yanfacheng123456@outlook.com

Traditional social media news is pure text-based news, which can be identified as fake through expert opinion, classification[2], graph models[3], and other methods. With the rapid development of multimedia and wireless communication technologies, the content of social media platforms is becoming increasingly diverse. Users can publish text, images, and short video messages on social media platforms, and they can easily fabricate, distort, and splice these messages to mislead readers. This brings new challenges to fake news detection.

This article aims to detect fake news that contains both text and images. Text and images provide rich information for detecting fake news, leading scholars to focus on the automatic detection of multimodal fake news. Currently, multimodal fake news detection methods mainly rely on the complementarity of text features and image features. For example, [4] attempted to learn a shared representation of text and images using an autoencoder to detect fake news. [5]utilized visual, textual, and social contextual information of news, and fused multimodal information based on an attention mechanism to detect fake news. [6] learned common features among all events in news and predicted the authenticity of news based on these features. [7] established a multimodal contextual attention network to fuse image and text features for fake news detection. Although these current methods for detecting fake news containing both images and text have shown good performance, there are still some drawbacks. Firstly, existing models often extract text and image features separately and concatenate them to exploit multimodal information, without considering the relationship between images and texts. This makes it difficult to learn high-order complementary information. Secondly, the components of existing detection models that capture the relationship between text and images are too simple to fully utilize textual and visual information to obtain high-order complementarity.

To address the aforementioned issues, this article proposes a multimodal fusion-based fake news detection model. The model uses a pre-trained VGG19[8] to learn the feature representation of news images, BERT[9] to learn the feature representation of news text, and a multimodal attention mechanism to fuse the intra-modality and inter-modality relationships, thereby capturing the high-order complementary relationship between text features and image features. Ultimately, the model learns an informative news representation for fake news detection.

The main contributions of this article are as follows:

A fake news detection model that fuses news text and images is proposed, which can effectively capture high-order complementary information between different modes for fake news detection.

We utilize a multimodal attention mechanism to model the features of news text and images, capturing effective information between the two modes in order to obtain an informative news representation.

Extensive experiments on real datasets have shown that our model is more effective and stable compared to the mainstream methods for fake news detection.

2 Related work

The key to multimodal fake news detection lies in how to comprehensively utilize the multimodal information of news to obtain an informative news representation. Summarizing existing articles, they can be roughly divided into two categories: multimodal fake news detection methods based on streaming and multimodal fake news detection methods based on graphing.

The detection methods based on the stream can be divided into single-stream architecture-based methods and multi-stream architecture-based methods. In single-stream architecture-based methods, multimodal data is fused through concatenation and function

mapping before being input into the model, and multimodal information needs to be learned in subsequent models. The main representatives are [10] and [11]. Multi-stream architecture-based methods design different models for each modality to learn the feature representation of each modality, jointly learn multimodal features from the feature representations of each modality, and input them into the downstream classifier to predict the authenticity of the news. The main representative is [5].

3 Problem statement

The key to fake news detection is to identify fake news on social media platforms, which is essentially a binary classification task. Given a news article N that contains both text and image information, the model will output $Y = \{0, 1\}$ to represent the news label. $Y = 0$ indicates that the news is a real news, while $Y = 1$ indicates that the news is fake.

4 Method

4.1 Model framework

As shown in Figure 1, our model consists of the following modules:

Information Encoding Module: To better capture the semantic meaning and contextual relationships of news text, we use BERT[9] to obtain a vector representation of the news text. For news images, we use a pre-trained model VGG19[8] to extract image features in a region-based manner. To avoid the problem of parameter explosion, we freeze the parameters of the pre-trained model during training.

Multimodal Attention Mechanism Module: We introduce a multimodal attention mechanism to fuse the information between news text and images. It can effectively integrate the internal relationships within the text modality and the relationships between the text and image modalities, obtaining informative news feature representations.

Classifier: The classifier utilizes a fully connected layer with an activation function to output predicted probabilities for each news item and determine whether the news is true or false.

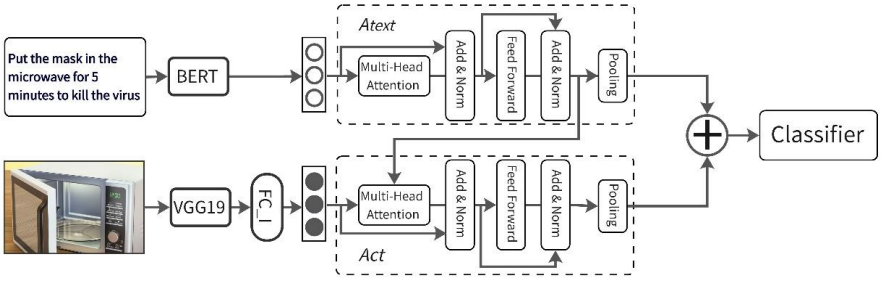

Fig. 1. The framework of a fake news detection model, where $\oplus$ represents the concatenation operation.

4.2 Information encoding module

We use $N = \{T, I\}$ to represent multimodal news that contains both textual and visual information, where T represents the textual information of the news, and I represents the visual information of the news.

Textual Encoder: We use BERT to capture the semantic meaning and contextual relationships of the news text. We first model the news text T as a sequence of words $T = \{t_1, t_2, \cdots, t_m\}$ (m being the number of words in the news text), and input the word sequence T into a pre-trained BERT to compute and obtain the feature representation $W = \{w1, w2, \cdots, w_m\}$ of the news text. The calculation formula is as follows:

$$W = \{w_1, \cdots, w_m\} = \mathrm{BERT}(T) \tag{1}$$

$w_i \in \mathbb{R}^{d_t}$ is the output layer hidden state of the corresponding token in BERT, and d_t is the dimension of the word embeddings.

Image Encoder. For the image information I in the news, we use the pre-trained model VGG-19 to extract the regional features of the image, and use the fully connected layer FC_I to obtain a feature representation with the same dimension as the text feature. The output set of regional features is represented by $P = \{p_1, p_2, ..., p_n\}$. During the training process, the parameters of VGG-19 are fixed. The calculation formula is as follows:

$$P = \{p_1, \cdots, p_m\} = Wvf[VGG-19(I)] \tag{2}$$

Where $p_i \in \mathbb{R}^{d_r}$ and d_r is the dimension of the image embedding. Wvf is the weight matrix of the fully connected layer.

4.3 Multimodal attention mechanism

As shown in Figure 1, the multimodal attention mechanism consists of a self-attention mechanism module (denoted as *Atext* in the figure) and a cross-modal attention mechanism module (denoted as *Act* in the figure).

The input of the self-attention mechanism module *Atext* is the news text feature W. The affinity matrix within the textual modality is calculated using the following formula:

$$A_w = softmax\left(\frac{FC_w^Q(W) \cdot FC_w^K(W)^{\mathsf{T}}}{\sqrt{d}}\right) \tag{3}$$

The $softmax$ operation is applied row-wise, while the fully connected layers FC_w^Q and FC_w^K represent different layers. Based on the affinity matrix within the textual modality, the text feature H_w can be represented as:

$$H_w^{'} = layer\_norm(W + A_w \cdot FC_w^V(W)) \tag{4}$$

$$H_w = layer\_norm(H_w^{'} + FC_w^{ff}(H_w^{'})) \tag{5}$$

The FC_w^V is a fully connected layer, and $layer\_norm$ is layer normalization. The FC_w^{ff} is a two-layer fully connected network, and the module *Atext* independently learns the text feature representation H_w without considering image information. To effectively fuse textual and visual information, we introduce a cross-modal attention mechanism

module *Act*, which further updates H_w with the aid of additional image features *P*. The process of calculating the cross-modal affinity matrix A_{co} is as follows:

$$A_{co} = softmax\left(\frac{FC_{co}^{Q}(P)\cdot FC_{co}^{K}(H_w)^{\mathsf{T}}}{\sqrt{d}}\right) \tag{6}$$

The affinity matrix $A_{co}[i,j]$ represents the importance of the *j* th word in the news text to the *i* th region in the news image. Then, *Act* uses the cross-modal affinity matrix A_{co} to learn the multimodal fused news text feature representation H_{co}, as shown below:

$$H_{co}^{'} = layer\_norm(P + A_{co}\cdot FC_{co}^{V}(H_w)) \tag{7}$$

$$H_{co} = layer\_norm(H_{co}^{'} + FC_{co}^{ff}(H_{co}^{'})) \tag{8}$$

The FC_{co}^{V} is a fully connected layer, and $layer\_norm$ is layer normalization. The FC_{co}^{ff} is a two-layer fully connected network. The next step is to perform pooling operations on H_w and H_{co}, and then concatenate the two types of features to obtain a multimodal fused news feature representation, denoted as H.

4.4 Classifier

Our classifier calculates the predicted probability of a news item through a fully connected layer with a $softmax$ activation function.

$$\hat{P}_i = \sigma(W_f H_i + b) \tag{9}$$

The *softmax* activation function is used in our classifier, where H_i represents the multimodal feature representation of the *i*th news item, $\hat{P}_i$ is the predicted probability of the *i*th news item, and W_f is the weight matrix of the fully connected layer. The cross-entropy loss function for our classifier is as follows:

$$\mathcal{L}(\Theta) = \sum_{i=1}^{N} -[\mathrm{Y}_i \log(\hat{P}_i) + (1-Y_i)\log(1-\hat{P}_i)] \tag{10}$$

N represents the number of news articles, and Y_i represents the predicted label for the *i*th news article.

5 Experiments

5.1 Datasets

We employed two real-world public datasets for our experiments: WEIBO [14] and TWITTER [4]. The WEIBO dataset contains post ID information, text information, and image information. The TWITTER dataset includes tweet text information, image

information, video information, and news social context information. For the WEIBO dataset, we followed the approach in [14] and divided 80% of the data as the training set and 20% as the testing set. For the TWITTER dataset, we used the development set for training and the testing set to evaluate the model performance. The statistical data for the two datasets is shown in Table 1.

Table 1. Statistics of the two datasets.

News	WEIBO	TWITTER
Fake News	4749	7898
Real News	4779	6026
Images	9528	514

5.2 Experimental details

We used Accuracy and F1 score as evaluation metrics in our experiments. The output dimension of the news text embedding from the BERT model is 768, while the regional feature dimension of the news image output from VGG-19 is 4096. We used a fully connected layer (FC_I) to transform the regional feature dimension from 4096 to 768, ensuring that the regional feature dimension and text embedding dimension are in the same feature space. The model in this paper is implemented based on the PyTorch learning framework [15], with a learning rate of 0.001, a training epoch of 130, a batch size of 128, and trained using the Adam[16] optimizer.

5.3 Baselines

We take the currently mainstream multi-modal fake news detection models as baseline models, specifically:

MVAE [4]: The fake news detection module uses the multimodal representation obtained from the double-peak variational autoencoder to classify posts as true or false.

EANN [6]: It uses an event discriminator to measure the differences between different events and further learn invariant features of events. By learning common features between different news articles, it detects fake news.

Att-RNN [14]: It proposes using an attention mechanism in an RNN to fuse multimodal features for fake news detection.

SAFE [17]: It jointly utilizes multimodal (textual and visual) and relational information to learn representations of news articles for fake news detection.

5.4 Results and analysis

As shown in Table 2, the proposed method in this paper outperforms all baseline methods on both datasets. On the WEIBO dataset, the detection accuracy of the proposed method is 5.6% higher than the baseline method, increasing from 81.9% to 87.5%, and the F1 score also increases from 80.3% to 86.9%. On the TWITTER dataset, there is a similar trend, with the accuracy increasing from 76.3% to 83.4%, and the F1 score increasing from 78.1% to 84.1%. The performance improvement is due to the proposed model's ability to capture comprehensive high-order complementary information between different modalities. In addition, the messages posted by users on social media platforms such as WEIBO and TWITTER are diverse, resulting in significant differences between the training and testing datasets for other models, which can impact their performance. Current models have not

effectively learned high-order complementary information within and across modalities, leading to poor model performance.

Table 2. Performance comparison of different models on two datasets.

Dataset	Methods	Accuracy	F1
WEIBO	SAFE	0.742	0.715
	Att-RNN	0.776	0.734
	EANN	0.780	0.755
	MVAE	0.819	0.803
	Our Model	0.875	0.869
TWITTER	SAFE	0.763	0.781
	Att-RNN	0.652	0.670
	EANN	0.639	0.607
	MVAE	0.719	0.742
	Our Model	0.834	0.841

6 Conclusions

This article proposes a multi-modal fusion fake news detection model that fully utilizes the multi-modal information of news content and effectively captures high-order complementary information within and across different modalities, enabling the model to learn rich and informative news feature representations for fake news detection. Experimental results demonstrate the proposed multi-modal fusion fake news detection model is more stable and effective compared to existing mainstream models. In future work, we will explore more effective methods for fake news detection, as an increasing number of users are posting content containing video information. Additionally, we will introduce social context information such as news comments, user information, and user forwarding relationships to further enhance the performance of fake news detection models.

References

1. SHU K, SLIVA A, WANG S, et al. Fake news detection on social media: a data mining perspective [J]. ACM SIGKDD Explorations Newsletter, 2017, 19(1): 22-36
2. WU K, YANG S, ZHU K Q. False rumors detection on Sinaweibo by propagation structures [C] // 2015 IEEE 31st International Conference on Data Engineering, Seoul, Korea, 2015: 651-662
3. GUPTA M, ZHAO P, HAN J. Evaluating event credibility on twitter[C] // Proceedings of the 2012 SIAM International Conference on Data Mining, Anaheim, USA, 2012: 153-164
4. Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: Multimodal variational autoencoder for fake news detection. In The World Wide Web Conference. 2915–2921
5. JIN Z, CAO J, GUO H, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs [C] // Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, USA, 2017: 795-816
6. WANG Y, MA F, JIN Z, et al. Eann: event adversarial neural networks for multi-modal fake news detection[C] // Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 2018: 849-857

7. Qian S, Wang J, Hu J, et al. Hierarchical multi-modal contextual attention network for fake news detection[C]//Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. 2021: 153-162.
8. Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).
9. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019.4171–4186.
10. KIM W, SON B, KIM I. ViLT: vision-and-language transformer without convolution or region supervision[C]//Proceedings of the 38th International Conference on Machine Learning, Jul 18-24, 2021: 5583-5594.
11. NAGRANI A, YANG S, ARNAB A, et al. Attention bottlenecks for multimodal fusion[C]//Advances in Neural Information Processing Systems 34, Dec 6-14, 2021: 14200-14213.
12. SONG C, SHU K, WU B. Temporally evolving graph neural network for fake news detection[J]. Information Processing & Management, 2021, 58(6): 102712.
13. CUI J, KIM K, NA S H, et al. Meta-path-based fake news detection leveraging multi-level social context information [C]//Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, Oct 17-21, 2022. New York: ACM, 2022: 325-334.
14. Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In Proceedings of the 25th ACM international conference on Multimedia. ACM, 795–816.
15. Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. (2017).
16. Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).
17. Xinyi Zhou, Jindi Wu, and R. Zafarani. 2020. SAFE: Similarity-Aware Multi-Modal Fake News Detection. ArXiv abs/2003.04981 (2020).