Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(8): 14699–14717.
DOI: 10.3934/mbe.2023657
Received: 20 April 2023
Revised: 15 June 2023
Accepted: 28 June 2023
Published: 06 July 2023

Research article

Multi-modal transformer for fake news detection

Pingping Yang[1], **Jiachen Ma**[1], **Yong Liu**[1,*] **and Meng Liu**[2]

[1] Heilongjiang University, Harbin 150000, China

[2] National University of Defense Technology, Changsha 410073, China

* **Correspondence:** Email: 2010023@hlju.edu.cn.

Abstract: Fake news has already become a severe problem on social media, with substantially more detrimental impacts on society than previously thought. Research on multi-modal fake news detection has substantial practical significance since online fake news that includes multimedia elements are more likely to mislead users and propagate widely than text-only fake news. However, the existing multi-modal fake news detection methods have the following problems: 1) Existing methods usually use traditional CNN models and their variants to extract image features, which cannot fully extract high-quality visual features. 2) Existing approaches usually adopt a simple concatenate approach to fuse inter-modal features, leading to unsatisfactory detection results. 3) Most fake news has large disparity in feature similarity between images and texts, yet existing models do not fully utilize this aspect. Thus, we propose a novel model (TGA) based on transformers and multi-modal fusion to address the above problems. Specifically, we extract text and image features by different transformers and fuse features by attention mechanisms. In addition, we utilize the degree of feature similarity between texts and images in the classifier to improve the performance of TGA. Experimental results on the public datasets show the effectiveness of TGA[*].

Keywords: fake news detection; multimodal fusion; attention mechanism; semantic matching

1. Introduction

As social networks expand their scope, more fake news emerges in online communities, which is detrimental to community stability and growth. On social media, fake news is a general information statement in some forms whose veracity is not quickly or ever confirmed [1]. This fake news frequently exists in the form of fake news in politics, economics, and public safety, which is tremendously hazardous to society. As an illustration, the COVID-19 outbreak resulted in the deaths of almost 800

[*] Our code is available at https://github.com/PPEXCEPED/TGA.

people due to the widespread misconception that consuming large amounts of alcohol might disinfect the body [2]. With the amount of online fake news on the rise, traditional manual methods can no longer handle the increasingly large volume of data. As a result, automated detection methods are gaining attention from academia and industry alike.

The traditional approaches are designed for text-only news. However, the prevalence of fake news with images on social media has sparked interest in methods that take multimodal inputs [3]. Many such methods have been proposed [3–7] in recent years, as fake news has evolved from text-only posts to multimedia posts with photos or videos [8]. Still, the following issues remain with current detection methods: 1) A lot of multimodal detection models [6, 9–11] currently extract visual features from news using pre-trained VGG-19 [12] on ImageNet [13] or ResNet-50 [14], which limits their ability to generate high-quality intermediate features and location information, resulting in unsatisfactory detection results. 2) Current multimodal fake news detection methods [5, 6, 15] essentially detect fake news by simply concatenating text and image features, ignoring the importance of different modalities to the news. 3) Existing multimodal fake news detection approaches [16–18] neglect the degree of feature similarity between multiple modalities despite considering the joint influence of different modal features.

To address the aforementioned challenges, we propose a novel multimodal framework called TGA that leverages transformer [19] to fully capture visual features, fuse multimodal features effectively, and make use of feature similarity between multi-modalities. Specifically, we use transformer and vision transformer to respectively extract text and image features, which are then fused by an attention mechanism to obtain the news representation. Finally, we put the news representation into our detector to detect rumors. To improve the performance of rumor detection, we also map the feature vectors of the two modalities to the same space for alignment and compute feature similarity between the two modalities. If the feature similarity between two modalities in a news is less than some threshold, we believe that the feature between two modalities in this news mismatch, so as to increase the probability that this news is Fake news. Thus we adjust the detection outcome of our detector according to the feature similarity between two modalities in a news, thereby improving detection performance. The main contributions of this paper are as follows:

- We introduce the TGA model, which utilizes various types of transformers to extract and represent visual and text features of news.
- We use attention mechanisms to fuse the representations of different modalities and calculate the degree of feature similarity between different modalities to obtain more robust representations and improve the performance of TGA.
- We evaluate the effectiveness of TGA on public datasets, demonstrating its superior performance compared to other state-of-the-art methods in detecting fake news.

2. Related work

In the field of fake news detection, existing methods mainly include three categories: 1) textual content-based, 2) visual content-based and 3) multimodal-based. In this section, we briefly review the work in recent years and explain the novelty of our method accordingly.

2.1. Textual content-based fake news detection

The textual content-based supervised fake news detection method uses textual content from the news as input to detect fake news. Ma et al. [20] first applied deep learning technology to fake news detection by feeding textual content into RNNs, LSTMs, and GRUs. Yu et al. [21] first used a convolutional neural network to model news. Ma et al. [22] applied the idea of multi-tasking for the first time, trained a multi-task model and position classification with the help of RNN. Ma et al. [23] used adversarial learning to detect fake news, improving the robustness and classification accuracy of the model. Vaibhav et al. [24] modeled article sentences as graphs, utilizing GCN to detect fake news and achieve positive results. Cheng et al. [25] used a variational autoencoder (VAE) to self-encode textual content to obtain an embedded representation of news and performed multi-task learning on the obtained news vectors to improve the model. [26] considering the temporal characteristics of rumors, this paper detects rumors using graph neural networks by leveraging the dynamic propagation structure of rumors. Moreover, many false news detection methods now utilize time temporal graphs [27] to construct graph structures. [28] used graph neural networks to extract text features, while utilizing user information and interaction information. However, previous works applied traditional RNN-based models to extract text features, which cannot be parallelized, and the physical meaning of feature extraction is unclear. In order to solve above problems, our work uses a transformer to extract text features, which not only achieves parallelization but also has a stranger explanatory model with its self-attention mechanism [19].

2.2. Visual content-based fake news detection

The news contains textual and visual content, such as images and videos. Recently, visual content has been demonstrated to be an essential indicator for detecting fake news [29, 30]. Traditional statistical-based methods detect fake news using the number of additional images, image popularity, and image type. With the rise of deep learning, many models use CNN, ResNet, and AlexNet to extract news features. However, traditional convolutional neural network models can only recognize pixel-level features of images. They cannot identify the semantic features of images, so they cannot detect whether images have been manipulated. Given that fake and real images can be very different in both the physical and semantic aspects, literature [31] proposes a fake image discriminator MVNN, which can effectively detect fake images. The best way to model image features has yet to be studied well in past studies, and most of it has focused on extracting text features. Many works use CNN-based models for image feature extraction, while our work introduces the vision transformer to fake news detection for the first time. Vision Transformer obtains global features from shallow layers, retains more spatial information than ResNet, and thus has a more vital ability to extract image features.

2.3. Multimodal-based fake news detection

Currently, more and more works consider using textual and visual content to detect fake news. With the rise of deep learning, many powerful feature extractors have emerged, such as text feature extractors RNN, Bert, and Transformer, and image feature extractors CNN, ResNet, and AlexNet. Text feature extractors can be used to extract text features, and image feature extractors can extract visual features, which are then fused for fake news detection.

Most of the works [6, 9, 10, 16, 27] directly concatenated the textual and image features obtained

from the extractor to detect fake news. For instance, literature [9] utilized VGG19 to extract visual content and XLNET to extract text content. Literature [10] used LSTM to model text content and text content in images, and used VGG to model visual content, Literature [6] used VGG to extract visual features and Text-CNN to extract visual features. Literature [16] extracts image features using VGG and text features via bi-directional LSTM.

Some work [27, 32] used the contrast between modalities to detect fake news. It has been asserted that news is fake if the visual and text content does not match. Based on this assumption, some people encoded the image and text information of the news and then calculated the similarity between the two. If the similarity is high, the news's text and visual information match and is real news. If the similarity is low, it means that the news's text and visual information do not match, and it is fake news. For example, literature [33] maps textual and visual information into the same vector space to compare the similarity to detect false news. Literature [32] uses BERT to model textual information and ResNet to model visual information to calculate the similarity between them. Inspired by the above-mentioned related works, we map the feature vectors into a new space to calculate similarity after obtaining the feature vectors of the two modalities.

There are also some works [17, 27, 34–38] that use multimodal information enhancement to detect fake news, where textual information can help to understand visual information and visual information can help to understand textual information. The mutual enhancement between the two modalities can be applied to detect fake news. For example, literature [17] first proposed using attention between modalities to enhance information between modalities; literature [27] employed the attention mechanism to obtain an enhanced visual representation of textual information to understand multimodal information better. Literature [35] designed a two-layer image-text co-attention to fuse visual information and textual information better, and literature [36] utilized the co-attention approach to learn more robust feature representations incorporating textual and visual information to enhance each other. However, these works ignore that different modalities have varying effects on fake news detection. Therefore, we should make the model pay attention to those significant modal information sources to improve its detection ability, according to [37] After using BERT to extract text features, this paper further utilizes BERT to extract both text and visual features, so as to enhance the mutual reinforcement between the two modal features. [38] use a cross-modal alignment module to transform the heterogeneous unimodality features into a shared semantic space. Inspired by the fusion of different modal features in the literature [17], our model uses an attention mechanism to stitch the two modal features together in late fusion.

3. Methodology

In this section, we will introduce the TGA model proposed in this paper. TGA is a transformer-based multimodal approach consisting of four key components: text feature extractor, image feature extractor, late fusion, and classifier. These components work together to extract and fuse text and image features, generating a comprehensive representation of the news that is then passed to the classifier for the task of rumor detection.

3.1. Model overview

The framework of TGA is illustrated in Figure 1. We start by obtaining the word embedding using Glove and then use transformer to generate the original vector set for the text. Next, we extract the original vector set for the image by processing different regions of the news image. We then pool the original vectors of text and image and employ an attention mechanism to fuse the guidance vectors of the two modalities, resulting in a final representation of the news. Finally, the news representation is fed into the MLP while mapping the guidance vectors of both modalities to a new target space to predict feature matching. The output of the MLP, combined with the feature similarity value weighted appropriately, obtains the final prediction result.

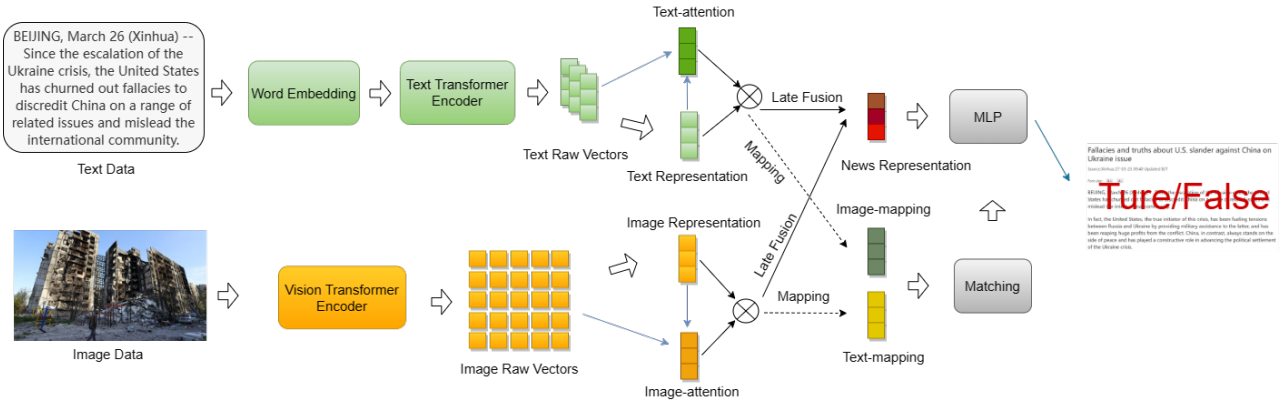

Figure 1. The overall framework of TGA. First, text features and visual features of each news are extracted by different types of transformers. Then, the features of two modal are fused by attention mechanism. Finally, multimodal feature similarity is added for further detection of fake news.

3.2. Text feature extractor

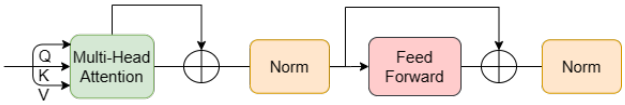

Figure 2. The structure of Transformer Encoder.

We obtain word embedding by the pre-trained model Glove [9] after utilizing the Jieba lexicon to segment the news texts. Given the transformer encoder's effectiveness in aggregating text features, we use it to extract text features. The word vectors obtained from the GloVe model are used as input for the transformer encoder. When encoding, we add position embeddings (PE) to the word vectors of each word. Specifically, we use sine and cosine position encoding, generated by applying sine and cosine functions of different frequencies to each position and then adding them to the corresponding word vectors. The calculation formula for PE is as follows:

$$PE(pos, 2i) = sin(\frac{pos}{1000^{\frac{2i}{d_{model}}}}) \tag{1}$$

$$PE(pos, 2i+1) = cos(\frac{pos}{1000^{\frac{2i}{d_{model}}}}) \tag{2}$$

where $PE \epsilon R^{L*d_{model}}$, L is the sentence length, which is 75 in this paper, d_{model} denotes the dimension size of the word vector, which is 512 in this paper, pos denotes the absolute position of the word in the sentence, and pos = 0, 1, 2,...,i indicates which dimension in the word vector. The input word vector is added to the position embedding, and the calculation formula is as follows:

$$X = GloveEmbedding(X) + PE \tag{3}$$

where $X \epsilon R^{L*d_{model}}$ denotes the word embedding of a news article, and GloveEmbedding is the operation to obtain the word embedding by the Glove model. After obtaining the word embedding from Eq (3), it is used as input for the transformer encoder. The transformer encoder comprises N block structures, as illustrated in the Figure 2. Each block consists of a multi-headed attention layer, residual connection layer, normalization layer, feedforward layer, residual connection layer, and normalization layer. In the first step, the calculation formula for word embedding in the multi-head attention layer is as follows:

$$Q = XW_Q, K = XW_K, V = XW_V \tag{4,5,6}$$

$$X_a = SelfAttention(Q, K, V) \tag{7}$$

$$SelfAttention(Q, K, V) = softmax(\frac{Q^T K}{\sqrt{d_k}})V \tag{8}$$

where W_Q,W_K,W_V is the matrix of three weights, d_k denotes the dimension of the matrix W_K, and Q^T is the transpose of Q. In the second step, take the residual connection of X_a obtained from Eq (7) is connected with X, then perform regularization, the calculation is as follows:

$$X_a = X + X_a \tag{9}$$

$$X_a = LayerNorm(X_a) \tag{10}$$

where LayerNorm denotes the regularization operation. In the third step, pass the regularized word embeddings to the input forward propagation layer. This layer consists of two linear connections and an activation function, and the calculation formula is as follows:

$$X_h = Activate(Linear(Linear(X_a))) \tag{11}$$

where *Activate* denotes the activation function, *Linear* denotes the fully connected layer. In the fourth step, the output of the forward propagation layer is then fed into the residual connection and regularization layer to obtain the final output $X_h \epsilon R^{L*d_{model}}$ of an encoding block, which is calculated as follows:

$$X_h = X_a + X_h \tag{12}$$

$$X_h = LayerNorm(X_h) \tag{13}$$

Equations (4)–(13) are repeated N times, which in the text, N=6. This paper refers to the hidden state vectors at different time points as the original vector set of the text. As mentioned earlier, the text guidance vector V_{text} is the result of pooling the original vector set of the text. Equation (14) shows the specific operation process.

$$V_{text} = \sum_{i=1}^{L} X_{hidden}^{i} \tag{14}$$

where L is the sentence length, which is set to 75 in this paper.

3.3. Image feature extractor

The supplied image is resized to 448*448 and sliced into 196 regions, each measuring 14*14 pixels. The regions are denoted by I_i (i = 1, 2,... 196). We employ ViT to fully extract the visual elements of news (Vision Transformer [39]). As a pre-trained model, ViT outperforms state-of-the-art image classification models on various image classification datasets and is relatively cost-effective. Moreover, when pre-trained on large-scale datasets and migrated to classification tasks on smaller and medium-sized datasets, ViT outperforms CNNs. [39] Therefore, we use the pre-trained ViT model to obtain the feature vector V_{region_i} of each region I_i, as shown in Eq (15). The ViT calculation process is the same as for the transformer encoder, and $N = 12$ is used in ViT. These region feature vectors are referred to as the original vector set of the image.

$$V_{region_i} = ViT(I_i) \tag{15}$$

As mentioned earlier, the image guidance vector is the result of pooling all the original vectors, as shown in Eq (16).

$$V_{image} = \frac{\sum_{i=1}^{N_r} V_{region_i}}{N_r} \tag{16}$$

where N_r is the number of regions, which is set to 196 in this paper.

3.4. Late fusion

To obtain final feature representation for news, we need to fuse the feature representation of different modality. Instead of simply concatenating the representations of different modalities, we employ an attention mechanism to fully integrate textual and visual representation into a multimodal representation. The attention mechanism has become a widely used component in deep learning to emphasize the most important information for the current task among several inputs, while ignoring insignificant information. To be specific, we compute the attention weights for each modality and create the final representation of the news via weighted averaging. To calculate the attention weights for modality m, we use a two-layer feedforward network with the following formula:

$$\widetilde{\alpha_m} = softmax(W_{m_2} \cdot tanh(W_{m_1} \cdot v_m + b_{m_1}) + b_{m_2}) \tag{17}$$

where $v_m \epsilon \left\{V_{text}, V_{image}\right\}$ represents the feature representation of modality m, $\widetilde{\alpha_m}$ represents the attention weight of modality m, W_{m_1}, W_{m_2} represents the weight matrix, and b_{m_1}, b_{m_2} represents the bias term. The feature presentation of modality m is then converted into a fixed-length form $v_m^{'}$ with the following formula:

$$v_m^{'} = tanh(W_{m_2} \cdot v_m + b_{m_2}) \tag{18}$$

The news feature representation v_f is then created by averaging and weighting the feature representations of all modes using the formula below:

$$v_f = \sum_{m \epsilon \{text, image\}} \widetilde{\alpha_m} v_m^{'} \tag{19}$$

3.5. Classifier

The classifier is a three-layer MLP that takes the news featue representation v_f obtained by late fusion as input for final classification. We denote the classifier as $G_r\left(v_f, \Theta_r\right)$, where Θ_r denotes all parameters in the classifier and the output of the classifier $\widetilde{y_f}$ is the probability that the news is fake news.

$$\widetilde{y_f} = G_r\left(v_f, \Theta_r\right) \tag{20}$$

The sigmoid activation function is utilized in the output layer to restrict the output values to 0 and 1. Through the examination of a significant amount of fake news detection data, we discovered that many fake news texts and images are not related. This is because many fake news writers use captivating images that have nothing to do with the text to attract readers. Therefore, we believe that computing the similarity of features across different modalities would enhance the detection of fake news due to the considerable differences in features between text and images found in such cases. To determine the degree of feature similarity, we map the feature representations of text and images to a new target space via calculation as follows:

$$S\left(V_{text}, V_{image}\right) = \left\| M_1\left(V_{text}\right) - M_2\left(V_{image}\right) \right\| \tag{21}$$

where S is the Euclidean distance of two modal features in the target space, $M_1\left(V_{text}\right)$ and $M_2\left(V_{image}\right)$ are two mapping functions, both of which consist of two layers of MLPs that map text and image feature representations to the new target space. We denote the final predicted values as:

$$\widetilde{y_f} = \begin{cases} \widetilde{y_f} + \alpha S\left(V_{text}, V_{image}\right) & \text{if } S\left(V_{text}, V_{image}\right) > \beta \\ \widetilde{y_f} & \text{if } S\left(V_{text}, V_{image}\right) \leq \beta \end{cases} \tag{22}$$

If the Euclidean distance between the two modalities is grater than the threshold β value, the result predicted by the classifier plus α times $S\left(V_{text}, V_{image}\right)$ is used as a reference. Where β and α are hyperparameters. The most effective parameter values we get through the experiments are $\beta = 0.65$ and $\alpha = 0.1$. If the final prediction value is greater than or equal to 0.5, we predict it as fake news, otherwise, we predict it as true news. Therefore, to calculate the classification loss, we use cross entropy, which is calculated as follows:

$$L_r\left(\Theta_r\right) = -y log\widetilde{y_f} - (1-y)\, log\left(1 - \widetilde{y_f}\right) \tag{23}$$

where y denotes the ground truth.

4. Experiments

In this section, we first introduce the dataset and parameter settings. Then we compare our proposed model TGA with several baselines and analyze the results of comparative experiments. Finally, we verify the effectiveness of each module of TGA by ablation study and dissect the impact of hyper-parameters by parameter sensitivity experiments.

Table 1. Dataset statistics.

	Domain/Statistics	Fake news	Real news	Total
Weibo	Finance	428	350	778
	Society	5642	5409	11,051
	Entertainment	556	733	1299
	Health	1756	1533	3289
Twitter	Training set	6840	5007	11,847
	Test set	564	427	991

4.1. Dataset and pre-treatment

4.1.1. Weibo dataset

The Weibo dataset utilized in this paper is retrieved from the DataFountain website (datafountain.cn). The multi-modal dataset is provided by the Beijing Municipal Bureau of Economy and Information Technology and the Big Data Expert Committee of the Chinese Computer Society and includes various fields such as Weibo texts, comments, images, and labels for three categories: "no judgment required", "fake news", and "real news". We selected only two labels: "fake news" and "real news". To clean up the dataset, we preserved only the Chinese characters of the Weibo text and removed content like emojis and meaningless symbols.

We also removed duplicate and low-quality images to ensure the dataset's quality. In this work, we focused on studying text and images, so text-only tweets were deleted, and only one image was kept for tweets with multiple images. After processing, 17,848 pieces of data totaled real and false news in eight categories: science and technology, politics, the military, finance and business, social life, sports and entertainment, medical and health, education, and examination. Due to a limited amount of data in the last four fields, we used data from the first four fields only. All the data in the first four categories were merged and randomly split into a training set (80%), a validation set (10%), and a test set (10%), totaling 16,417 items. Table 1 displays the dataset's specifics.

4.1.2. Twitter

The Twitter [40] dataset was released for Verifying Multimedia Use task at MediaEval. In experiments, we keep the same data split scheme as the benchmark [40]. The training set contains 6840 real tweets and 5007 fake tweets, and the test set contains 991 posts, including 564 real tweets and 427 fake tweets. In experiments, we follow the same steps in weibo dataset to remove the duplicated and low-quality images to ensure the quality of the entire dataset.

4.2. Parameter settings

The text feature extractor and image feature extractor produce output dimensions of 256 and 1024, respectively. The mapping function generates output dimensions of 128 for both text and image features. Furthermore, the text transformer implements multi-headed attention with eight heads, while the image transformer utilizes 16 heads. During training, we employ a batch size of 32, a learning rate of 0.001, and optimize the loss function using the Adam optimizer. To achieve faster convergence, we use

a dynamic learning rate method. We record the F1-Score after each epoch and adjust the learning rate to 80% of the previous epoch's rate if the F1-Score does not improve from the previous epoch. Finally, we evaluate model performance using precision, recall, accuracy, and F1-Score.

4.3. Baselines

To verify the effectiveness of our multimodal model, we compare it with the following baselines:

Unimodal Models

- **CNN** [41]: A CNN-based model which uses CNN to extract image features and employs a three-layer neural network for classification.
- **LSTM** [42]: A textual model which using LSTM to extract text features of news.

Multimodal Models

- **EANN** [6]: A model uses a CNN-based extractor to extract text features and a VGG-19 network to extract image features.
- **MVAE** [4]: A model extracts text and image features of news and reconstructs the original image and text from the hidden layer vectors.
- **Spotfake+** [9]: A multimodal model that utilizes transfer learning to capture semantic and contextual information from news texts and their associated images.
- **Att-RNN** [17]: Combine textual, visual, and social contextual features by attention mechanism.
- **MCAN** [35]: An end-to-end model which using multiple co-attention layers to fuse image and text features, which can learn the interdependencies between multiple modalities.
- **HMCAN** [43]: Model multimodal features of news by a multimodal contextual attention network so that information from different modalities complements each other.

4.4. Comparative experiment

Table 2 shows the performance of baselines and our model; we can obtain the following points from the experimental results:

- Multi-modal models perform significantly better than Unimodal models, which indicates the effectiveness of detecting fake news using multi-modal information.
- Spotfake+ outperforms att-RNN while utilizing pre-trained feature extractors for feature extraction because pre-training typically improves a model's capabilities for generalization and expedites its convergence to the target task.
- HMCAN is superior to Spotfake+ after modality augmentation with a contextual attention network, it indicates the effectiveness of the attention mechanism in fake news detection.
- We can observe that on both datasets, the performance of MCAN is noticeably better than HMCAN. Because MCAN uses two feature extractors to fully extract image features not only highlights the significance of attention mechanisms in multi-modal fusion but also emphasizes the massive contribution of image features to rumor detection.
- Our proposed model TGA outperforms the best baseline model MCAN, although MCAN uses co-attention in multimodal fusion, it ignores the importance of the degree of feature similarity between different modalities for rumor detection, so MCAN does not detect as well as our model

TGA. This not only further proves that our feature extractor is superior to traditional CNN and traditional RNN-based feature extractors but also illustrates the significant role of multimodal feature similarity in rumor detection.

For a more visual representation of the comparison experiment results, we plotted the line graphs depicted in Figures 3 and 4, where the horizontal axis shows the comparison models and the vertical axis represents the values of the four evaluation metrics.

Table 2. Comparative experiments.

	Model	Precision	Recall	Accuracy	F1-Score
Weibo	CNN	0.592	0.806	0.556	0.683
	LSTM	0.757	0.590	0.647	0.663
	EANN	0.925	0.736	0.740	0.820
	MVAE	0.931	0.708	0.723	0.804
	SpotFake+	0.871	0.871	0.870	0.871
	att-RNN	0.741	0.777	0.772	0.723
	MCAN	0.899	**0.899**	0.902	0.900
	HMCAN	0.888	0.885	0.885	0.885
	TGA	**0.969**	0.886	**0.922**	**0.925**
Twitter	CNN	0.452	0.539	0.425	0.479
	LSTM	0.554	0.431	0.511	0.523
	EANN	0.745	0.748	0.745	0.744
	MVAE	0.697	0.627	0.688	0.639
	att-RNN	0.691	0.692	0.662	0.682
	MCAN	0841	0.847	0.889	0.849
	HMCAN	0.876	**0.888**	0.878	0.875
	TGA	**0.912**	0.854	**0.918**	**0.918**

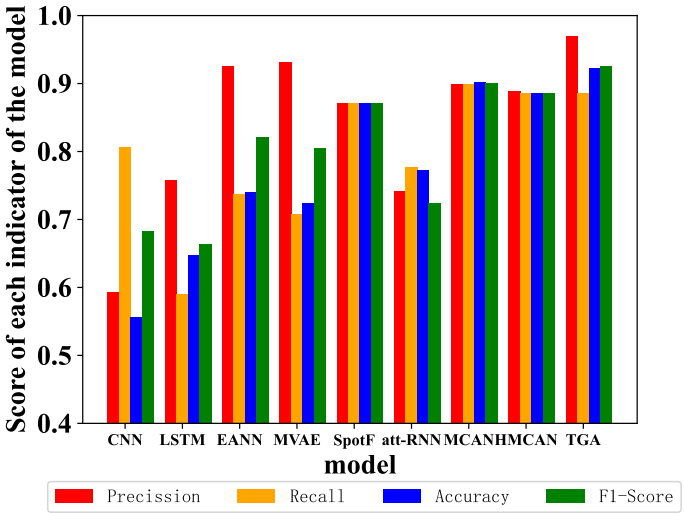

Figure 3. Comparison of the four assessment results of the experiment (Weibo).

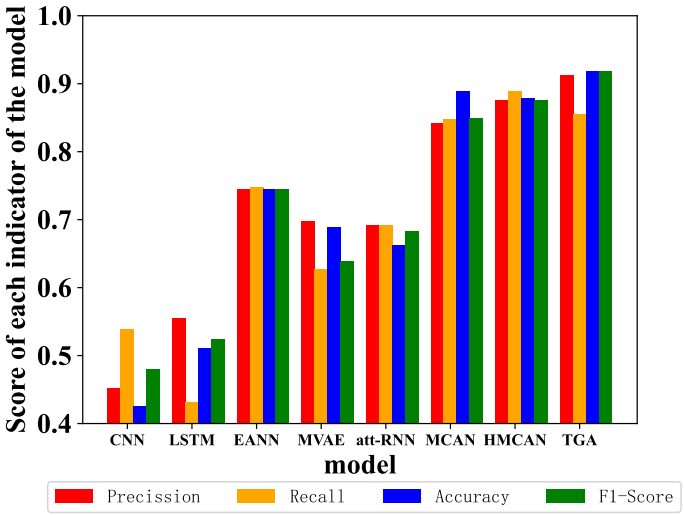

Figure 4. Comparison of the four assessment results of the experiment (Twitter).

4.5. Ablation study

Table 3. Ablation experiments.

	Model	Precision	Recall	Accuracy	F1-Score
Weibo	TGA-T	0.669	0.670	0.660	0.664
	TGA-I	0.895	0.867	0.880	0.880
	TGA-A	0.914	**0.914**	0.906	0.914
	TGA-L	0.905	0.878	0.891	0.892
	TGA-R	0.790	0.823	0.906	0.806
	TGA-M	0.886	0.859	0.867	0.872
	TGA	**0.969**	0.886	**0.922**	**0.925**
Twitter	TGA-T	0.548	0.557	0.548	0.550
	TGA-I	0.745	0.769	0.772	0.776
	TGA-A	0.887	**0.891**	0.884	0.896
	TGA-L	0.896	0.847	0.897	0.902
	TGA-R	0.735	0.796	0.870	0.756
	TGA-M	0.842	0.793	0.814	0.857
	TGA	**0.912**	0.854	**0.918**	**0.918**

In order to verify the effectiveness of each module of TGA, we compare each of the following variants with TGA:

- **TGA-T**: Only text is used, the image feature part is deleted.
- **TGA-I**: Only the image is used, the text feature part is deleted.
- **TGA-A**: The part based on attention mechanism fusion is removed and directly concatenates the features of the two modalities.
- **TGA-L**: The transformer is replaced with LSTM in the text feature extractor.

- **TGA-R**: The VIT is replaced with ResNet-50 in the image feature extractor.
- **TGA-M**: The impact of feature similarity calculation results is removed from the experiment.

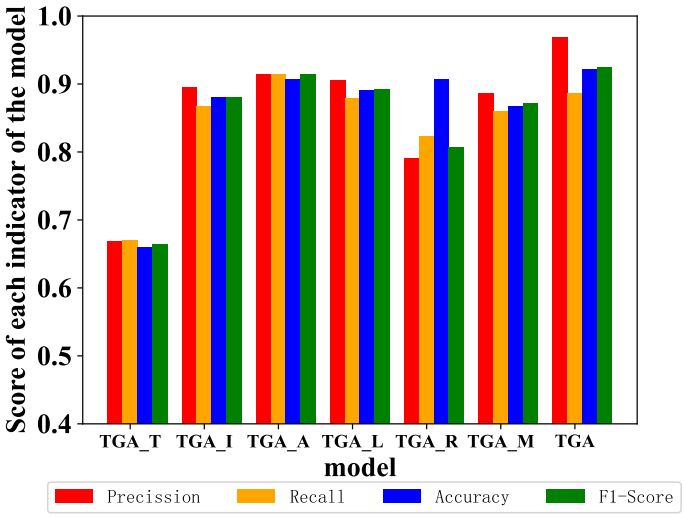

Figure 5. Results of ablation experiments on four benchmarks (Weibo).

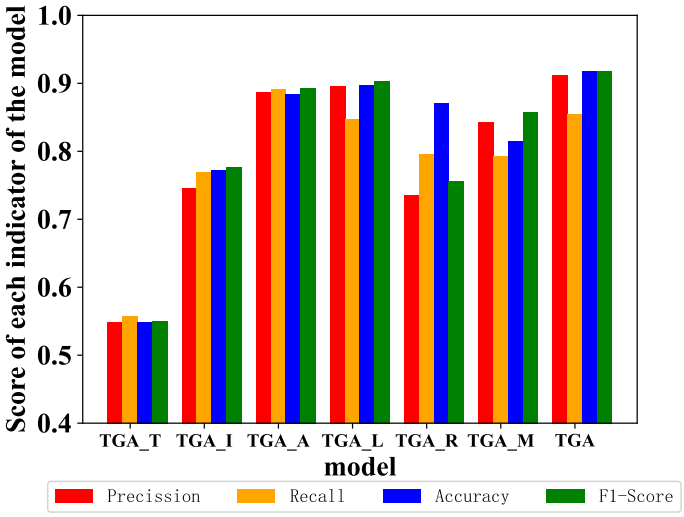

Figure 6. Results of ablation experiments on four benchmarks (Twitter).

Table 3 shows the experimental results of several variants and we can obtain the following points:

- TGA outperforms all variants, which indicates the effectiveness of each module of TGA.
- TGA-T and TGA-I have the worst performance among all variants proving that multimodal detection is superior to unimodal.
- TGA-I is superior to TGA-T, which illustrates the image-based modality model is more effective than the text-based modality model. This is because that it is difficult to distinguish between true and false news according to the text content as they usually contain many similar field-specific terms. However, fake news is often artificially created by using images unrelated to the content to

attract attention. When the images contained in a news do not match the field to which the news belongs, the news will easily be identified as Fake news. For this reason, image-based detection is often more effective than text-based detection when dealing with fake news in the same field.

- TGA outperforms TGA-A which indicates that the effectiveness of attention mechanism in multimodal fusion. Due to attention mechanism can help model find the most inportant information.
- TGA-L is inferior to TGA, indicating that Transformer is better than the traditional RNN-based feature extractor for extracting text features. Similarly, TGA-R is inferior to TGA, which proves that VIT is better than the traditional CNN-based feature extractor in extracting image features.
- TGA outperforms TGA-M, because multimodal feature similarity provides the degree of matching between modalities to enhance the model's capability, also demonstrating that the level of semantic matching between multiple modalities significantly impacts news detection.

Additionally, we utilized bar charts, as presented in Figures 5 and 6, to illustrate the results of the ablation experiment in a clearer manner.

4.6. Parameter sensitivity experiments

The results of our experiments are highly sensitive to the chosen hyperparameters. To provide insights into their effects on the experimental outcomes, we showcase selected hyperparameter results in Figures 7–9. Notably, we conducted all hyperparameter experiments exclusively on the Weibo dataset.

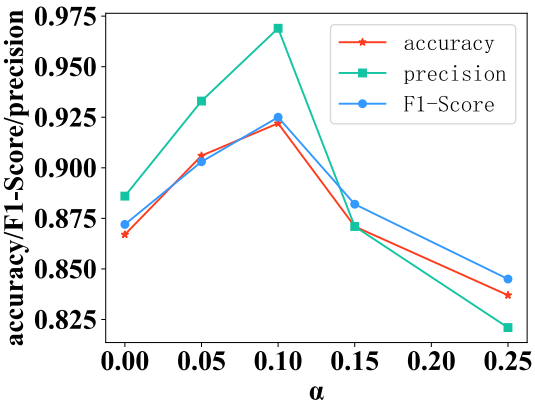

Figure 7. Effect of α on model performance.

Figure 7 shows the impact of the threshold α value on the experimental results. The fraction of the feature similarity degree of the two modal features in the experimental results is measured using a threshold α value. The final prediction result is calculated using the classifier's prediction result plus α times the feature similarity value. The experimental results show that the more significant α, the greater the influence of the feature similarity degree. Setting α to 0.1 permits us to achieve the best performance for the model.

Figure 8 shows the impact of the threshold β value on the experiment results. Our experiments report that the optimal performance is achieved when β is set to 0.65. When the feature similarity value of the two modalities outweighs β. In that case, We believe that there is a significant disparity in the similarity between features from the two modalities, and we will add the feature similarity value

of α times the feature similarity value to the classifier prediction result to evaluate whether the news is fake.

In Figure 9, we illustrate the impact of the word embedding dimension on our experimental outcomes. We observed that a word embedding dimension of 32 yields the best results for our model. Our analysis suggests that when the word embedding dimension is below 32, the vector representation of the words is insufficient to capture word features accurately. As we increase the word embedding dimension beyond 32, the language's inherent ambiguity amplifies, leading to overfitting.

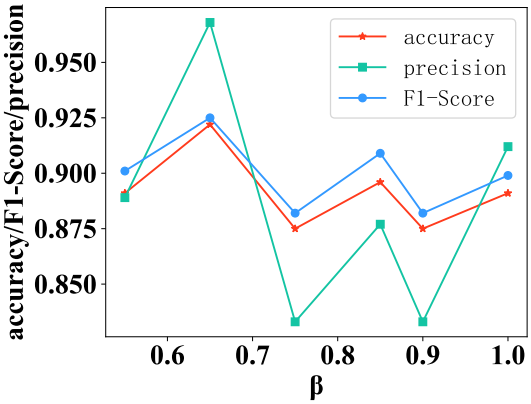

Figure 8. Effect of β on model performance.

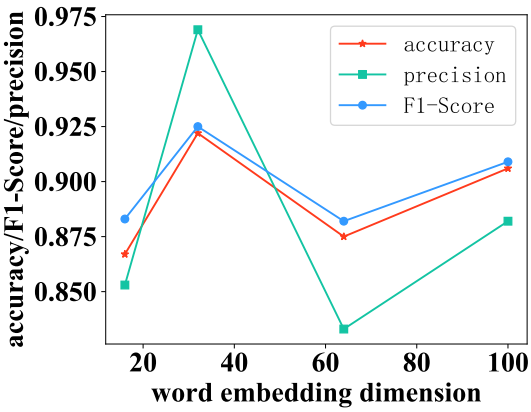

Figure 9. Effect of word embedding dimension on model performance.

5. Conclusions

In this paper, we propose a transformer-based multi-modal model TGA to study the problem of detecting multi-modal fake news. Specifically, we use a different type of transformer to extract textual and image features and employ attention mechanisms to fuse multi-modal features in the late stage. In addition, we calculate the semantic matching degree of multiple features to improve the detection effect. Experimental results on real datasets show that our proposed model outperforms existing multi-modal models. We will consider improving the TGA for cross-domain news detection in future work.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the Natural Science Foundation of Heilongjiang Province in China (No. LH2020F043).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. S. M. Alzanin, A. M. Azmi, Detecting rumors in social media: A survey, *Procedia Comput. Sci.*, **142** (2018), 294–300. https://doi.org/10.1016/j.procs.2018.10.495
2. S. Islam, T. Sarkar, S. H. Khan, A. Kamal, H. Seale, A. Kabir, et al., COVID-19-related infodemic and its impact on public health: A global social media analysis, *Am. J. Trop. Med. Hyg.*, **103** (2020), 1–9. https://doi.org/10.1038/s41598-020-73510-5
3. Z. W. Jin, J. Cao, G. Han, Y. D. Zhang, J. B. Luo, Multimodal fusion with recurrent neural networks for rumor detection on microblogs, in *Proceedings of the 25th ACM International Conference on Multimedia*, (2017), 759–816. https://doi.org/10.1145/3123266.3123454
4. D. Khattar, J. S. Goud, M. Gupta, V. Varma, MVAE: Multimodal variational autoencoder for fake news detection, in *The World Wide Web Conference*, (2019), 2915–2921. https://doi.org/10.1145/3308558.3313552
5. S. Singhal, A. Kabra, M. Sharma, R. R. Shah, P. Kumaraguru, SpotFake: A multi-modal framework for fake news detection, in *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, (2019), 39–47. https://doi.org/10.1109/BigMM.2019.00-44
6. Y. Q. Wang, F. L. Ma, Z. W. Jin, Y. Yuan, G. X. Xun, K. Jha, et al., EANN: Event adversarial neural networks for multi-modal fake news detection, in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2018), 849–857. https://doi.org/10.1145/3219819.3219903
7. H. W. Zhang, Q. Fang, S. S. Qian, C. S. Xv, Multi-modal knowledge-aware event memory network for social media rumor detection, in *Proceedings of the 27th ACM International Conference on Multimedia*, (2019), 1942–1951. https://doi.org/10.1145/3343031.3350850
8. J. Cao, P. Qi, Q. Sheng, T. Y. Yang, Exploring the role of visual content in fake news detection, *Disinf. Misinf. Fake News Social Media*, **2020** (2020), 141–161. https://doi.org/10.1007/978-3-030-42699-6_8
9. S. Singhal, A. Kabra, M. Sharma, R. Shah, P. Kumaraguru, SpotFake+: A multimodal framework for fake news detection via transfer learning (student abstract), in *Proceedings of the AAAI Conference on Artificial Intelligence*, **34** (2020), 13915–13916. https://doi.org/10.1609/aaai.v34i10.7230

10. J. S. Liu, K. Feng, J. Z. Pan, J. Deng, L. Wang, MSRD: Multimodal web rumor detection method, *J. Comput. Res. Dev.*, **11** (2020), 9. https://doi.org/10.21203/rs.3.rs-101168/v1

11. T. Jin, H. X. Xia, Lookback option pricing models based on the uncertain fractional-order differential equation with Caputo type, *J. Ambient Intell. Hum. Comput.*, **2021** (2021), 1–14. https://doi.org/10.1007/s12652-021-03516-y

12. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, preprint, arXiv:1409.1556.

13. D. Jia, D. Wei, R. Socher, L. J. Li, L. Kai, F. F. Li, ImageNet: A large-scale Hierarchical Image Database, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, (2009), 248–255. https://doi.org/10.1109/CVPR.2009.5206848

14. K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90

15. Y. Wang, F. Ma, H. Wang, K. Jha, J. Gao, Multimodal emergent fake news detection via meta neural process networks, in *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2021), 3708–3716. https://doi.org/10.1145/3447548.3467153

16. D. Khattar, J. S. Goud, M. Gupta, V. Varma, MVAE: Multimodal variational autoencoder for fake news detection, in *the World Wide Web Conference*, (2019), 2915–2921. https://doi.org/10.1145/3308558.3313552

17. Z. Jin, J. Cao, G. Han, Y. Zhang, J. Luo, Multimodal fusion with recurrent neural networks for rumor detection on microblogs, in *Proceedings of the 25th ACM International Conference on Multimedia*, (2017), 795–816. https://doi.org/10.1145/3123266.3123454

18. M. Liu, Z. W. Quan, J. M. Wu, Y. Liu, M. Han, Embedding temporal networks inductively via mining neighborhood and community influences, *Appl. Intell.*, **2022** (2022), 1–20. https://doi.org/10.1007/s10489-021-03102-x

19. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, (2017), 6000–6010. https://doi.org/10.5555/3295222.3295349

20. J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K. Wong, et al., Detecting rumors from microblogs with recurrent neural networks, in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, (2016), 3818–3824.

21. F. Feng, Q. Liu, S. Wu, L. Wang, T. Tan, A convolutional approach for misinformation identification, in *Twenty-Sixth International Joint Conference on Artificial Intelligence*, (2017), 3901–3907. https://doi.org/10.5555/3172077.3172434

22. J. Ma, W. Gao, K. F. Wong, Detect rumor and stance jointly by neural multi-task learning, in *Companion Proceedings of the Web Conference 2018*, (2018), 585–593. https://doi.org/10.1145/3184558.3188729

23. J. Ma, W. Gao, K. F. Wong, Detect rumors on twitter by promoting information campaigns with generative adversarial learning, in *The World Wide Web Conference*, (2019), 3049–3055. https://doi.org/10.1145/3308558.3313741
24. V. Vaibhav, R. Mandyam, E. Hovy, Do sentence interactions matter? leveraging sentence level representations for fake news classification, in *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing*, (2019), 134–139. https://doi.org/10.18653/v1/d19-5316
25. M. X. Cheng, S. Nazarian, P. Bogdan, VRoC: Variational autoencoder-aided multi-task rumor classifier based on text, in *Proceedings of the Web Conference 2020*, (2020), 2892–2898. https://doi.org/10.1145/3366423.3380054
26. C. G. Song, K. Shu, B. Wu, Temporally evolving graph neural network for fake news detection, *Inf. Process. Manage.*, **58** (2021), 102712. https://doi.org/10.1016/j.ipm.2021.102712
27. M. Liu, K. Liang, B. Xiao, S. H. Zhou, W. X. Tu, Y. Liu, et al., Self-supervised temporal graph learning with temporal and structural intensity alignment, preprint, arXiv:2302.07491. https://doi.org/10.48550/arXiv.2302.07491
28. Y. Q. Jin, X. T. Wang, R. C. Yang, Y. Z. Sun, W. Wang, H. Liao, et al., Towards fine-grained reasoning for fake news detection, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **36** (2022), 5746–5754. https://doi.org/10.48550/arXiv.2110.15064
29. M. X. Cheng, S. Nazarian, P. Bogdan, Fake news detection on social media: A data mining perspective, *ACM SIGKDD Explor. Newsl.*, **19** (2017), 22-36. https://doi.org/10.1145/3137597.3137600
30. Z. W. Jin, J. Cao, Y. D. Zhang, J. S. Zhou, Q. Tian, Novel visual and statistical image features for microblogs news verification, *IEEE Trans. Multimedia*, **19** (2019), 598–608. https://doi.org/10.1109/TMM.2016.2617078
31. P. Qi, J. Cao, T. Y. Yang, J. B. Guo, J. T. Li, Exploiting multi-domain visual information for fake news detection, in *2019 IEEE International Conference on Data Mining*, (2019), 518–527. https://doi.org/10.1109/ICDM.2019.00062
32. J. Xue, Y. Wang, Y. Tian, Y. Li, L. Wei, Detecting fake news by exploring the consistency of multimodal data, *Inf. Process. Manage.*, **58** (2021), 102610. https://10.1016/j.ipm.2021.102610
33. X. Zhou, J. Wu, R. Zafarani, SAFE: Similarity-aware multi-modal fake news detection, preprint, arXiv:2003.04981.
34. H. Zhang, Q. Fang, S. Qian, C. Xu, Multi-modal knowledge-aware event memory network for social media rumor detection, in *the 27th ACM International Conference*, (2019), 1942–1951. https://doi.org/10.1145/3343031.3350850
35. Y. Wu, P. Zhan, Y. Zhang, L. Wang, Z. Xu, Multimodal fusion with co-attention networks for fake news detection, in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, (2021), 2560–2569. https://doi.org/10.18653/v1/2021.findings-acl.226
36. W. Zhang, L. Gui, Y. He, Supervised contrastive learning for multimodal unreliable news detection in COVID-19 pandemic, in *the 30th ACM International Conference on Information and Knowledge Management*, (2021), 3637–3641. https://doi.org/10.1145/3459637.3482196

37. J. H. Hua, X. D. Cui, X. H. Li, K. K. Tang, P. C. Zhu, Multimodal fake news detection through data augmentation-based contrastive learning, *Appl. Soft Comput.*, **136** (2023), 1568–4946. https://doi.org/10.1016/j.asoc.2023.110125

38. Y. X. Chen, D. S. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, et al., Crossmodal ambiguity learning for multimodal fake news detection, in *Proceedings of the ACM Web Conference 2022*, (2022), 2897–2905. https://doi.org/10.1145/3485447.3511968

39. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. H. Zhai, T. Unterthiner, An image is worth 16x16 words: Transformers for image recognition at scale, in *International Conference on Learning Representations*, 2021. https://doi.org/10.48550/arXiv.2010.11929

40. C. Boididou, S. Papadopoulos, M. Zampoglou, L. Apostolidis, Y. Kompatsiaris, Detection and visualization of misleading content on Twitter, *Int. J. Multimedia Inf. Retr.*, **7** (2017), 71–86. https://doi.org/10.1007/s13735-017-0143-x

41. Y. Lecun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, et al., Backpropagation applied to handwritten zip code recognition, *Neural Comput.*, **1** (1989), 541–551. https://doi.org/10.1162/neco.1989.1.4.541

42. P. Zhou, W. Shi, J. Tian, Z. Y. Qi, B. C. Li, H. W. Hao, et al., Attention-based bidirectional long short-term memory networks for relation classification, in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, **2** (2016), 207–212. https://doi.org/10.18653/v1/P16-2034

43. Z. Jin, J. Cao, G. Han, Y. D. Zhang, J. B. Luo, Multimodal fusion with recurrent neural networks for rumor detection on microblogs, in *Proceedings of the 25th ACM International Conference on Multimedia*, (2017), 795–816. https://doi.org/10.1145/3123266.3123454

AIMS Press