Multimodal false information detection method based on Text-CNN and SE module

Yi Liang¹,², Turdi Tohti¹,²*, Askar Hamdulla¹,²

¹ School of Information Science and Engineering, Xinjiang University, Urumqi, China, ² Xinjiang Key Laboratory of Signal Detection and Processing, Urumqi, China

* turdy@xju.edu.cn

Abstract

False information detection can detect false information in social media and reduce its negative impact on society. With the development of multimedia, the multimodal content contained in false information is increasing, so it is important to use multimodal features to detect false information. This paper mainly uses information from two modalities, text and image. The features extracted by the backbone network are not further processed in the previous work, and the problems of noise and information loss in the process of fusing multimodal features are ignored. This paper proposes a false information detection method based on Text-CNN and SE modules. We use Text-CNN to process the text and image features extracted by BERT and Swin-transformer to enhance the quality of the features. In addition, we use the modified SE module to fuse text and image features and reduce the noise in the fusion process. Meanwhile, we draw on the idea of residual networks to reduce information loss in the fusion process by concatenating the original features with the fused features. Our model improves accuracy by 6.5% and 2.0% on the Weibo dataset and Twitter dataset compared to the attention based multimodal factorized bilinear pooling. The comparative experimental results show that the proposed model can improve the accuracy of false information detection. The results of ablation experiments further demonstrate the effectiveness of each module in our model.

Introduction

With the development of information technology, social media has become the main way for people to obtain information, especially during the epidemic, people’s lives are more closely connected with social media. The rapid development of social media not only brings convenience to people, but also facilitates the spread of false news and misleading news. False information is defined as unsubstantiated stories and statements [1]. The spread of false information can mislead the public and have a negative impact on society. For example, the 2012 Doomsday Theory, which declares that the Earth will experience a major catastrophe on December 21, 2012, or “three consecutive days of darkness”, this rumor has caused panic among people around the world, causing people to spend a lot of money to hoard shopping supplies, and even spend a lot of money to build “Noah’s Ark”.

Fig 1 presents several multimodal false information posts from the Twitter dataset [2]. Each post contains a paragraph of text and an image associated with it. The image in the first post has been altered so the image is false and the text is false, the second post image is real but the image is about the Sicily air disaster. In the final post, the image was edited to include a shark that didn’t exist during Hurricane Sandy. The dissemination of this false information has a serious impact on the normal operation of society, so it is very important to find out how to detect false information and stop it from spreading. For false information containing text and images, can be divided into three categories. The first type of false information is that the text content is false but the image is true, the second type of false information is that the text content is true but the image is false, and the third type of false information is that both text content and images are false.

In recent years, deep learning models have been used for false information detection, early approaches focused on detection using text features [3, 4], such as the model proposed by Pérez [5] uses textual content to detect false information, but this model can only detect the first and third types of false information, but cannot correctly detect the second type of false information, if we use both text and image information at the same time, all kinds of false information can be detected [6–8], which reflects the importance of multi-modal false information detection.

There are two challenging problems in existing research work. First, how to extract higher-quality text features and image features. Second, how to better fuse text features and image features to obtain more valuable fusion features. Previous works have used RNN (Recurrent Neural Network) [9] or Transformer-based models [10] to extract text features, CNN (Convolutional Neural Network)-based [11] models to extract image features, and finally fusing text and image features through simple splicing, factorization bilinear pool or attention mechanism. However, these methods directly fuse the features extracted from the backbone network, and fail to perform corresponding processing on the extracted features to make up for the insufficiency of the features extracted from the backbone network in some aspects. Furthermore, these research works do not consider the problems of noise and information loss in the feature fusion process. This paper proposes a false information detection model based on Text-CNN [12] and SE (Squeeze-and-Excitation Networks) [13] modules, which solves the above problems well. Our model uses 3 scales of Text-CNN to make up for the slight deficiency.
of the Transformer-based model in the extraction of local features, at the same time, we adopt our modified SE module to reduce the influence of noise in the fusion process, and reduce the information loss in the fusion process by concatenating the original features and the fusion features. Therefore, our model can better detect false information.

The main contributions of this paper are as follows:

1. We use 3 different scales of Text-CNN to process the text features extracted by the pre-trained model BERT [14] and the image features extracted by the pre-trained model SWTR [15] to obtain more valuable features.

2. We modify the SE module so that it can fuse text features and image features. We utilize the channel attention mechanism in the SE module to mitigate the effect of noise during fusion to obtain better-represented fused features.

3. We draw on the idea of residual network to concatenate original features and fusion features to reduce the information loss in the fusion process.

4. The accuracy and F1 value of our model on Twitter dataset and Weibo dataset [16] outperform the baseline model AMFB.

Related work

Traditional false information detection models are mostly text-based. In earlier studies, people mainly extracted text features manually. Qazvinian et al. [17] exploited n-grams, bi-grams features extracted from text to detect rumors. Pérez et al. [5] extracted five linguistic features from the text to detect error messages. With the development of technology, researchers found that the artificial extraction of features will be limited by the dataset, resulting in the extracted features not having generality [18]. Subsequent researchers used deep learning techniques to allow computers to automatically extract features from text to detect false information. Liu et al. [19] proposed a model that uses CNN to extract text features and detect false information. The model uses CNN to mine deeper text features that humans cannot discover. Ma et al. [20] used RNN to extract text features, and the model used RNN to discover content related to textual contextual content. Nasir [21] utilizes both CNN and RNN to extract text features, which can combine the advantages of CNN and RNN.

In recent years, with the increasing number of forms of information expression, how to use different forms of information simultaneously to detect false information detection has attracted the attention of many researchers. Two forms of information, text and images, are often used in existing research work to detect false information. Singhal et al. [22]. The text and image features extracted by BERT and VGG19 are concatenated and fed into the classifier to obtain detection results. Kumari et al. [23]. proposed a multi-modal fusion model based on multi-modal factorization bilinear pooling, the model first uses a combination of BILSTM and Attention to extract text features, followed by a combination of CNN, BiGRU and Attention to extract image features, and finally the two features are fused by Multimodal Factorized Bilinear Pooling and fed into the detector to obtain detection results. Song et al. [24]. proposed a multi-modal false information detection model based on cross-modal attention residuals and multi-channel CNN. The model can extract information related to the target modality from the remaining modalities without losing the information of the target modality, while the influence of noise during the fusion of information from different modalities can be reduced by multi-channel CNN. Dhawan [25] proposed a multi-modality detection model based on a graph neural network, which can allow fine-grained interactions within and between different modalities to further improve the accuracy of multi-modal false information detection.
et al. [26], proposed a novel multimodal co-attention network to better fuse text and image features for false information detection. With the rise of pre-trained models, researchers have conducted research on fusion algorithms that fuse text features and image features. Xu et al. [27] divided the existing Transformer—based pre-trained fusion models into six categories, (1) Early Summation [28, 29], the model takes text and image features, weights them together and feeds them into the Transformer layer to fuse text and image features. This fusion method does not increase the computational complexity, but requires manual setting of the weights. (2) Early Concatenation [30–33], this model concatenates text features and image features and then inputs them into the Transformer layer to fuse the features of different modalities. This approach increases computational complexity. (3) Multi-stream to One-stream [34], this model inputs text features and image features into two Transformer layers for processing, and then concatenates them through another Transformer layer to fuse features. (4) One-stream to Multi-stream [35], this model concatenates text features and image features and inputs them into the Transformer layer for fusion, and then divides the fused features into two parts and inputs them into two different Transformer layers. (5) Cross-Attention [36, 37], using two Transformer layers to process text features and image features by exchanging two Q(Query) to complete the fusion of text features and image features. (6) Cross-Attention to Concatenation [38, 39], the text features and image features processed by Cross-Attention are concatenated and input to another Transformer layer for processing.

In addition to text and image information, other forms of information can also be used to detect false information. Wang [40] and others found that the existing research ignores the role of the strong emotion of the image in the rumor content, proposes a multimodal rumor detection model composed of visual emotion and textual emotion. Azri [41] proposed an end-to-end model that utilizes three features of text, images and emotion simultaneously. Armin [42] proposed a multimodal detection model that supports the fusion of different levels and types of information, which can simultaneously utilize textual, visual, user reviews and metadata.

Multimodal false information detection method based on Text-CNN and SE module

Problem Definition: Suppose $P = \{p_1, p_2, \ldots, p_m\}$ each post in this dataset contains data in the form of both text and images, where $p_i$ represents the $i$th post. $T_{set} = \{t_1, t_2, \ldots, t_m\}$ is the text set, $t_i$ represents the text content in the $i$th post, $V_{set} = \{v_1, v_2, \ldots, v_m\}$ is the image set, $v_i$ represents the image contained in the $i$th post, $L = \{l_1, l_2, \ldots, l_m\}$ is the tag set, $l_i$ is the tag of the $i$th post, $p_i = \{t_i, v_i, l_i\}$. The main purpose of disinformation detection is to find a function $f(T, V) = Y$, this function identifies the authenticity of a post by the text information and image information in the post. $Y = \{y_1, y_2, \ldots, y_m\}, y_i$ is the predicted label of the $i$th post.

The model in this paper mainly consists of four parts: text feature extraction, image feature extraction, image and text feature fusion and classifier. Fig 2 shows our proposed multimodal false information detection method based on Text-CNN and SE module.

The model first uses BERT to extract the features of each token from the text, and concatenates the features of each token as text features, and its dimension is (33/95,768). Use SWTR to extract image features with dimension (49,768), then fuse text features and image features through modified SE module to obtain fused features, and text and image features are processed by Text-CNN with widths of 1, 2 and 3 to improve feature quality. Then the fused features, locally enhanced text features and image features are concatenated and fed into a classifier to classify them.
Text feature extraction

Usually, when people post on social media, they express their thoughts in the form of text. The text contains the main meaning that the publisher wants to express. Therefore, how to process the text and extract high-quality text features has a significant impact on the detection accuracy of the model. This paper extracts text features through the combination of BERT and Text-CNN models.

BERT is a Transformer-based pre-training model. BERT is first trained on a large unsupervised dataset to learn some general knowledge, and then the learned knowledge is transferred to a specific task. BERT, due to its special structure, can achieve good results while reducing the consumption of training resources. However, the extraction of local features by BERT is slightly inadequate, so this paper uses Text-CNN to process the text features extracted by BERT, so that the text features contain more local information. Text-CNN is a CNN model applied to text. The convolution kernel of Text-CNN will keep the length of the convolution kernel consistent with the length of the word feature, and only adjust the width of the convolution kernel. The model can extract similar features to n-grams. The calculation of text features extracted using BERT is as follows:

\[ PE = \text{Position} \_ \text{Embeddings}(t) \]  
\[ SE = \text{Segment} \_ \text{Embeddings}(t) \]  
\[ TE = \text{Token} \_ \text{Embeddings}(t) \]  
\[ T_{\text{all}} = PE + SE + TE \]

\( \text{Position} \_ \text{Embeddings}() \) is a function that encodes the position of the text, \( \text{Segment} \_ \text{Embeddings}() \) is a function that encodes the paragraphs of the text, and \( \text{Token} \_ \text{Embeddings}() \) is a function that transforms each word in the text into a word vector, where \( t \) is a set of text, \( T_{\text{all}} = \{ T_{[d]} \}, T \) is the text feature extracted by BERT, \( T \in R^{nd}, T_{[d]} \in R^{d}, n \) is the number of words in a sentence, and \( d \) is the dimension of each token feature vector. In this paper, we process \( T \) by three
scales of Text-CNN, and the calculation formula is as follows:

\[
T_1 = \phi(\text{conv}1(W_1, T) + b_1)_{k=1}
\]

\[
T_2 = \phi(\text{conv}1(W_2, T) + b_2)_{k=2}
\]

\[
T_3 = \phi(\text{conv}1(W_3, T) + b_3)_{k=3}
\]

\(\phi\) is the activation function, \(\text{conv}1()\) is the function of the one-dimensional convolution operation, \(W_1, W_2, W_3\) are convolution kernels that can be obtained by learning, and \(b_1, b_2, b_3\) are bias that can be obtained by learning, and \(k\) represents the width of the convolution kernel. \(T_1, T_2, T_3 \in \mathbb{R}^{64}\) are the text features obtained after three widths of Text-CNN processing.

### Image feature extraction

Images are more believable than text content, so an accurate image feature extraction module plays an important role in false information detection models. We use SWTR to extract image features and further process the extracted image features through Text-CNN with three widths.

SWTR is a successful model for using Transformer in computer vision. SWTR can extract both local features and global features through a windowing mechanism compared to CNN-based models, and SWTR’s unique windowing mechanism reduces computational effort compared to the rest of Transformer-based models. SWTR has SOTA performance on multiple tasks. Since SWTR and BERT are both Transformer-based models, this model is therefore also similar to the BERT model in that it has some shortcomings in the treatment of local features. Since the windowing mechanism of SWTR cannot set the size of the convolutional kernel as flexibly as CNN, this paper uses Text-CNN to process the image features extracted by SWTR. In this paper, the image features extracted by SWTR are input into three different widths of Text-CNN for processing. The specific calculation process is as follows:

\[
V = \text{SWTR}(v)
\]

\[
V_1 = \phi(\text{conv}1(W_4, V) + b_4)_{k=1}
\]

\[
V_2 = \phi(\text{conv}1(W_5, V) + b_5)_{k=2}
\]

\[
V_3 = \phi(\text{conv}1(W_6, V) + b_6)_{k=3}
\]

\(\text{SWTR}()\) is the function of extracting image features using Swin-Transformer, \(\phi\) is the activation function, \(\text{conv}1()\) is the function for the 1D convolution operation, \(W_4, W_5, W_6\) are the learnable convolution kernels, \(b_4, b_5, b_6\) are the biases, \(k\) is the scale of the convolution kernel, \(v\) represents an image in the post, \(V \in \mathbb{R}^{z\times dv}\) is the image feature, \(z\) is the number of extracted features, and \(dv\) is the dimension of the feature vector. \(V_1, V_2, V_3 \in \mathbb{R}^{64}\) are the image features after processing by Text-CNN of three widths.

### Feature fusion

So far we have the text features \(T_{\text{all}}\) extracted by BERT, the image features \(V\) extracted by SWTR, text features \(T_1, T_2, T_3\) and image features \(V_1, V_2, V_3\). The SE module is mostly used for channel feature enhancement of the input feature maps in computer vision tasks. For example, if we input a feature map \(A\) with dimensions \((H, W, C)\), the SE module will input \(A\) into two
full connection layers to get the attention score, and then multiply the feature map \( A \) with the attention score in the dimension of the channel to get the final output. We modify the SE module so that it can fuse text features and image features to obtain multimodal fusion features. The MSE module is shown in Fig 3. The SE module can assign weights to each channel.

![Diagram of SE module](https://doi.org/10.1371/journal.pone.0277463.g003)
automatically filter low-weight noise points. With a small increase in the number of parameters, the performance of the model on related tasks can be greatly increased. Therefore, we modify the SE module to fuse text and image features. The calculation process is as follows:

\[
\begin{align*}
    \text{Score}_{\text{test}} &= \text{Linear}(V) \\
    T' &= \text{Bmm}(\text{Score}_{\text{test}}, T) \\
    T'' &= \text{avgpooling}(\beta(T')) \\
    \text{Score}_{\text{image}} &= \text{Linear}(T) \\
    V' &= \text{Bmm}(\text{Score}_{\text{image}}, V) \\
    V'' &= \text{avgpooling}(\beta(V'))
\end{align*}
\] (12-17)

In this paper, the attention scores between different modalities are obtained through the fully connected layer. \(\text{Score}_{\text{image}} \in \mathbb{R}^{n\times z}, \text{Score}_{\text{test}} \in \mathbb{R}^{n} \) are the attention scores of text features on image features and the attention scores of image features on text features. \(\text{Bmm}()\) is the function that performs the dot product operation. \(T' \in \mathbb{R}^{n}, V' \in \mathbb{R}^{z}\) are image features and text features fused by the MSE module, \(\beta\) is the activation function. Then we concatenate the text features and image features processed by Text-CNN and the fusion features extracted by the MSE module. The calculation process is as follows:

\[
    F = \text{concatenate}(T_1, T_2, T_3, V_1, V_2, V_3, T'', V'')
\] (18)

\(\text{Concatenate}()\) is a function of concatenation operation, \(F \in \mathbb{R}^{64+64+64} \) is the fusion feature that is finally input into the classifier.

**False information detection**

We feed the fused features into fully connected layer and Softmax layer to obtain detection results.

\[
\begin{align*}
    F' &= \text{Linear}(F) \\
    p_i &= \text{Softmax}(F') \\
    y_i &= \text{arg max}(p_i)
\end{align*}
\] (19-21)

We use the cross-entropy loss function to calculate the loss value:

\[
    L = -\sum_{i=1}^{m} [l_i \log p_i + (1 - l_i)\log(1 - p_i)]
\] (22)

\(p_i\) is the probability that the post is false, \(\text{argmax}\) is the function of select the predicted label values, \(y_i\) is the predicted label value of the post by the model, \(m\) is the number of posts, and \(l_i \in \{0, 1\}\) is the true label value, 1 represents false information, and 0 represents true information.

**Experiment and analysis**

**Dataset and experimental settings**

Machine configuration and environment for this experiment: CPU: Intel Xeon E5-2630L v3, 62 G memory, 8 cores, GPU: NVIDIA GeForce RTX 3090, PyTorch(1.7.1), Python(3.8), Cuda
To compare with previous work, we use the Twitter and Weibo datasets to complete our experiments. These are two publicly available, high quality datasets that can be used for multimodal disinformation detection.

The Twitter dataset is published by Boididou et al. The dataset contains training dataset and test dataset. The training dataset contains three types of information: false, true and humorous, but the test dataset contains only two types of information: true and false, so we remove the humorous type of information from the training dataset. The Weibo dataset is a multimodal Chinese dataset that contains only two types of posts, real and false. We split the posts containing multiple images in the Twitter dataset and the Weibo dataset into multiple posts containing only one image, and deleted the data containing only images, only text and images as gifs and black and white images. Table 1 shows the data distribution for the Weibo and Twitter datasets.

The two datasets above are publicly available datasets applied to false information detection studies. We are only interested in the text, images and labels in the dataset, so some remaining information is removed. We take text and images as input to the model and labels as facts. First we preprocess the text and images, for the text part we remove punctuation, URL and emoticons in the sentence, and for the image part, we resize all the images to (224, 224, 3). The training dataset is used to train the model, and the test dataset is used to verify the performance of the model.

Table 2 lists all the hyperparameters used to train the model.

Comparative experiment

We implement some uni-modal and multimodal models to verify the validity of our model.

- **BERT**: We use $T_{[cls]}$ extracted from the fine-tuned BERT-Base as text features. The text features $T_{[cls]}$ are fed into the classifier to detect the authenticity of posts.

- **SWTR**: We use the image feature $V$ extracted from the SWTR model to feed into the average pooling layer to obtain the image features $V_m$, and then input the processed image feature into the classifier to detect the authenticity of posts.

Table 1. Data distribution of the Weibo and Twitter datasets.

| Dataset | Train | Test | Image |
|---------|-------|------|-------|
|         | False | True | False | True  |       |
| Twitter | 6827  | 4993 | 717   | 1215  | 410   |
| Weibo   | 6476  | 4096 | 1136  | 1215  | 13274 |

Table 2. Hyperparameters used in model training.

| Parameters | Twitter | Weibo |
|------------|---------|-------|
| Text length| 33      | 95    |
| Image size | (224,224,3) | (224,224,3) |
| Batch size | 70      | 70    |
| Optimizer  | Adam(lr = 0.0001) | Adam(lr = 0.00005) |
| Epochs     | 100     | 100   |
| Dropout    | 0.2     | 0.6   |
Multimodal based models:

- **att-RNN** [16]: att-RNN is an RNN with attention mechanism that fuses text and image features for false information detection.

- **EANN** [43]: EANN (Even Adversarial Neural Network, EANN) is an end-to-end event adversarial network that uses an event discriminator to remove the impact of event information on detection results and improve the generality of the model.

- **MVAE** [44]: MVAE (Multimodal Variational Autoencoder, MVAE) is used to learn the correlation between modalities and then combined with a classifier to detect false information.

- **AMFB** [23]: AMFB (Attention based multimodal Factorized Bilinear, AMFB), the network uses BILSTM and VGG19 to extract text and image features, finally uses multimodal decomposition bilinear pooling to fuse features of text and images.

In order to verify the effectiveness of our proposed model, we compare the above baseline model with our model on both Weibo and Twitter datasets. At the same time, we conduct 5 experiments for each model under the same experimental conditions and take the average of the 5 experiments as the final result, the aim of which is to reduce the influence of experimental errors on the experimental results. The experimental results are shown in Table 3. According to the accuracy and F1 value, our model has better performance than the existing baseline models. It can also be observed that the performance of the single-modal model is lower than the multimodal model on both datasets, this suggests that using both text and image information can be more effective in detecting false information.

Figs 4 and 5 show the accuracy and loss values of our model when trained on the Twitter and Weibo datasets. The full form of ‘iter’ is ‘iterations’. As can be seen from the figure, the loss gradually decreases to an equilibrium position, followed by a slight fluctuation at the equilibrium position, which indicates that the model is learning properly. It can be observed from the figure that our model is fully trained on both datasets, while our model has more difficulty in achieving convergence on the Weibo dataset because the Weibo dataset contains more images and different posts from different events, while most of the posts on the Twitter dataset are from the same event.

### Table 3. Comparative results for the Weibo and Twitter datasets.

| Dataset | Model | Accuracy | F1 value of False News | F1 value of Real News |
|---------|-------|----------|------------------------|-----------------------|
| Twitter | BERT  | 0.831    | 0.852                  | 0.802                 |
|         | SWTR  | 0.834    | 0.871                  | 0.766                 |
|         | att-RNN | 0.664   | 0.676                  | 0.651                 |
|         | EANN  | 0.741    | 0.610                  | 0.810                 |
|         | MVAE  | 0.745    | 0.758                  | 0.730                 |
|         | AMFB  | 0.883    | 0.920                  | 0.810                 |
|         | OUR   | 0.903    | 0.924                  | 0.866                 |
| Weibo   | BERT  | 0.870    | 0.880                  | 0.858                 |
|         | SWTR  | 0.713    | 0.702                  | 0.724                 |
|         | att-RNN | 0.772   | 0.692                  | 0.754                 |
|         | EANN  | 0.791    | 0.780                  | 0.80                  |
|         | MVAE  | 0.824    | 0.809                  | 0.837                 |
|         | AMFB  | 0.832    | 0.840                  | 0.830                 |
|         | OUR   | 0.897    | 0.902                  | 0.890                 |

[https://doi.org/10.1371/journal.pone.0277463.t003](https://doi.org/10.1371/journal.pone.0277463.t003)
Ablation experiment

We set up 4 ablation experiments to demonstrate the effectiveness of our model.

- Ablation Experiment 1: We compare our model with the original model after removing different modules to demonstrate the validity of each module.
- Ablation Experiment 2: We compare the original BERT and SWTR models with our improved BERT and SWTR models to demonstrate the validity of our improvements.
- Ablation Experiment 3: We set up a series of experiments to demonstrate that the model is most effective in processing text and image features using three different scales of Text-CNNs simultaneously.
- Ablation Experiment 4: Several different fusion methods were used to demonstrate the effectiveness of the fusion methods we used by comparing them with the fusion methods we used.
For the above 4 groups of ablation experiments, in order to eliminate errors, we performed 5 experiments for each model and took the average value.

**Ablation experiment one.** To demonstrate the effectiveness of each module in our proposed model, we conduct ablation experiments and the results are shown in Table 4:

- **OUR**: The complete model presented in this paper.
- **-SE**: We simply splice the text features $T_1$, $T_2$, $T_3$ extracted by BERTcnn and the image features $V_1$, $V_2$, $V_3$ extracted by SWTRcnn to detect the authenticity of the post. But remove MSE module.
- **-Text-CNN**: We use the MSE module to fuse text features $T$ and image features $V$, excluding text features and image features processed by Text-CNN.
- **-SE-Text-CNN**: We simply concatenate the text feature $T_{[cls]}$ extracted by BERT-Base and the image feature $V_a$ processed by the average pooling layer and input it into the classifier to detect the authenticity of the post. Removed the MSE module and Text-CNN module.

From Table 4, the complete model achieves the best results, demonstrating the effectiveness of each module. We can observe that the model with any module removed shows a decrease in accuracy compared to OUR. -SE dropped by 0.4%, -Text-CNN dropped by 0.8%, -SE-Text-CNN dropped by 1.0%. The MSE module in the model can alleviate the problem of noise introduced in the fusion process, so that the model can better fuse text and image features. It also reduces information loss during fusion by simply concatenating text features and image features that have been processed by Text-CNN.

**Ablation experiment two.** To verify that our improvements to BERT and SWTR are effective. We compare the original BERT and SWTR with our improved BERT and SWTR, and the results are shown in Fig 6.

- **BERT**: We input the text features $T_{[cls]}$ extracted by BERT into the classifier to get the detection result of the post.
- **SWTR**: We input the image features $V_a$ obtained from the averaging pooling layer into the classifier to obtain the classification results of the post.
- **BERT$_{cnn}$**: We obtained features $T_1$, $T_2$ and $T_3$ by processing the text features $T$ extracted by BERT with three different scales of Text-CNN. Subsequently, $T_1$, $T_2$ and $T_3$ are concatenated and fed into the classifier to obtain detection results.
- **SWTR$_{cnn}$**: We use three different scales of Text-CNN to process the image features $V$ to obtain features $V_1$, $V_2$ and $V_3$, which are subsequently concatenated and fed into the classifier to obtain detection results.

As can be seen in Table 4, our improvements to BERT and SWTR are effective. The BERTcnn model improved the accuracy of the Weibo dataset by 1.1% compared with the
BERT, and the SWTRcnn model improved the accuracy of the Weibo dataset by 1.6% compared with the SWTR model. The experimental results prove our conjecture that the text and image features extracted by the Transformer-based pre-trained model can be further improved by Text-CNN processing.

Ablation experiment three. In order to demonstrate the effectiveness of using three different scales of Text-CNN to process the features extracted by BERT and SWTR, we compare with the following models, and the results are shown in Figs 7 and 8.

- **BERTcnn**: We use Text-CNN with 64 convolution kernels of size (1,768) to process the text feature $T$ extracted by BERT-Base, and input the processed text features into the classifier to obtain classification results.

- **BERTcnn2**: We use Text-CNN with 64 convolution kernels of size (1,768) and 64 convolution kernels of size (2,768) to process the text feature $T$, and concatenating it input into the classifier to obtain detection results.

- **BERTcnn3**: We use Text-CNN with 64 convolutional kernels of size (1,768), 64 convolutional kernels of size (2,768), 64 convolutional kernels of size (3,768) and 64 convolutional kernels
of size (4,768) to process the text features $T$ and concatenate them into the classifier to obtain the results.

- **BERT$_{cnn}$**: The model used in this paper for detecting the authenticity of posts after processing the text features $T$ using Text-CNN at three different scales.
- **SWTR$_{cnn1}$**: Similar to the model $BERT_{cnn1}$, except that the processed text features $T$ are replaced with image features $V$.
- **SWTR$_{cnn2}$**: Similar to the model $SWTR_{cnn2}$, just replace the processed text features $T$ with image features $V$.
- **SWTR$_{cnn4}$**: Replace the text feature $T$ processed by $SWTR_{cnn4}$ with the image feature $V$.
- **SWTR$_{cnn}$**: The image features $V_1$, $V_2$ and $V_3$ are concatenated and fed into the classifier to obtain the detection results.

As can be seen in Figs 7 and 8, the accuracy of $BERT_{cnn1}$, $BERT_{cnn2}$, $BERT_{cnn}$ and $SWTR_{cnn1}$, $SWTR_{cnn2}$, $SWTR_{cnn}$ for post detection gradually improves as the number of Text-CNNs of different scales increases. However, when the number of convolution kernels at different scales
exceeds three. The performance of the model will gradually decrease. Comparing $\text{BERT}_{\text{cnn}}$, $\text{SWTR}_{\text{cnn}}$ with $\text{BERT}_{\text{cnn4}}$, $\text{SWTR}_{\text{cnn4}}$, the accuracy increases by 0.3% and 0.8%, and the F1 value also increases by 0.3% and 0.6%. The experimental results show that our proposed Text-CNN with three scales (1,768), (2,768) and (3,768) are the most effective to process text features and image features.

We analyze the dataset to further validate our conclusions. We used the jieba word splitter to segment the test dataset from the Weibo dataset and to calculate the number of tokens of different lengths. The distribution is shown in Fig 9.

From Fig 9, we can see that the length of each token in the sentence is not consistent, so when we use Text-CNN of different scales to extract local features similar to n-grams in the text, we will not only extract some valid features, at the same time, some invalid features will be extracted. From Fig 9 we can see that 97% of the tokens in the dataset are less than 4 in length and only 3% of tokens are longer than 3, combined with the experimental results in Fig 7, it can be concluded that when we use Text-CNN with widths of 1, 2, and 3 to extract features, more valid features are extracted than invalid features, thus increasing the performance of the model. When we use larger width Text-CNN, more invalid features than valid features are extracted, thus degrading the performance of the model.
Ablation experiment four. To demonstrate the effectiveness of our fusion method, we set up several different fusion models to compare with our fusion model.

- E-Sum (Early Summation): The features of the different modalities are weighted and summed by position and fed into the Transformer for processing.

- E-Con (Early Concatenation): The features of the different modalities are concatenated and fed into the Transformer for processing.

- M-to-O (multi-stream to one-stream): First, two Transformer layers are used to process text features and image features, and then concatenate and input into another Transformer layer for processing.

Fig 9. Frequency of words with different lengths.
https://doi.org/10.1371/journal.pone.0277463.g009
• O-to-M (one-stream to multi-stream): First, the text features and image features are concatenated and input into the Transformer for processing, and then split and input into two Transformers layers for processing.

• Cross-A (Cross-Attention): When using two Transformer layers to process text features and image features, exchange two Q (Query) to complete the fusion of text features and image features.

• Cross-A-C (Cross-Attention to Concatenation): The text features and image features processed by Cross-A are concatenated and input to another Transformer layer for processing.

As can be seen in Fig 10 our model has the best performance, which proves the effectiveness of the feature fusion method we use, and the fusion method we use has a smaller number of parameters than other fusion methods.

As shown in the Table 5, the number of parameters for the fusion methods we used is much smaller than for the rest of the fusion methods.
Conclusion

This paper proposes a multimodal false information detection method based on Text-CNN and SE module. The model first uses multi-scale Text-CNN to process text features and image features, and uses the MSE module to fuse multi-modal features to obtain fusion features. Finally, the text features and image features processed by Text-CNN and the fusion feature is simply concatenated as the final fusion feature to detect false information. The comparative experiments demonstrate that our model achieves better results on the Weibo and Twitter datasets than the rest of the models. The ablation experiments validate the effectiveness of our improvements to each module of the model.

In future work, we will mainly study the following issues: (1) How to reduce the size of the model so that it can be deployed on small devices while ensuring detection accuracy. (2) How to extract higher quality features from text and images (3) How to fuse text features and image features more fully.

Author Contributions

Conceptualization: Turdi Tohti.
Investigation: Turdi Tohti.
Methodology: Yi Liang.
Project administration: Askar Hamdulla.
Software: Yi Liang.
Supervision: Turdi Tohti, Askar Hamdulla.
Validation: Yi Liang, Turdi Tohti, Askar Hamdulla.
Visualization: Yi Liang.
Writing – original draft: Yi Liang.
Writing – review & editing: Turdi Tohti, Askar Hamdulla.

References

1. Gupta, Manish and Zhao, Peixiang and Han, Jiawei. Evaluating event credibility on twitter. Proceedings of the 2012 SIAM international conference on data mining. Society for Industrial and Applied Mathematics, 2012: 153-164.
2. Boididou C, Andreadou K, Papadopoulos S, et al. Verifying multimedia use at mediaeval. MediaEval, 2015, 3(3): 7.
3. Rashkin H, Choi E, Jang J Y, et al. Truth of varying shades: Analyzing language in fake news and political fact-checking. Proceedings of the 2017 conference on empirical methods in natural language processing. 2017: 2931-2937.

4. Popat K, Mukherjee S, Strötgen J, et al. Credibility assessment of textual claims on the web. Proceedings of the 25th ACM international conference on information and knowledge management. 2016: 2173-2178.

5. Pérez-Rosas V, Kleinberg B, Lefevre A, et al. Automatic detection of fake news. arXiv preprint arXiv:1708.07104, 2017.

6. Alonso-Bartolome S, Segura-Bedmar I. Multimodal Fake News Detection. arXiv preprint arXiv:2112.04831, 2021.

7. Peng X, Xintong B. An effective strategy for multi-modal fake news detection. Multimedia Tools and Applications, 2022, 81(10): 13799–13822. https://doi.org/10.1007/s11042-022-12290-8

8. Choi H, Ko Y. Effective fake news video detection using domain knowledge and multimodal data fusion on youtube. Pattern Recognition Letters, 2022, 154: 44–52. https://doi.org/10.1016/j.patrec.2022.01.007

9. Graves A. Long short-term memory. Supervised sequence labelling with recurrent neural networks, 2012: 37–45. https://doi.org/10.1007/978-3-642-24797-2_4

10. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in neural information processing systems, 2017, 30.

11. Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84–90. https://doi.org/10.1145/3065386

12. Chen Y. Convolutional neural network for sentence classification. University of Waterloo, 2015.

13. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

14. Lee J D C K, Toutanova K. Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

15. Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.

16. Jin Z, Cao J, Guo H, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. Proceedings of the 25th ACM international conference on Multimedia. 2017: 795-816.

17. Qazvinian V, Rosengren E, Radev D, et al. Rumor has it: Identifying misinformation in microblogs. Proceedings of the 2011 conference on empirical methods in natural language processing, 2011: 1589-1599.

18. MENG J, WANG L, YANG Y, et al. Multi-modal deep fusion for false information detection. Journal of Computer Applications, 2022, 42(2): 419.

19. LIU Z, WEI Z, ZHANG R. Rumor detection based on convolutional neural network. Journal of Computer Applications, 2017, 37(11): 3053.

20. Ma J, Gao W, Mitra P, et al. Detecting rumors from microblogs with recurrent neural networks. 2016.

21. Nasir J A, Khan O S, Varlamis I. Fake news detection: A hybrid CNN-RNN based deep learning approach. International Journal of Information Management Data Insights, 2021, 1(1): 100007. https://doi.org/10.1016/j.jmi.2020.100007

22. Singhal S, Shah R R, Chakraborty T, et al. Spotfake: A multi-modal framework for fake news detection. 2019 IEEE fifth international conference on multimedia big data (BigMM). IEEE, 2019: 39-47.

23. Kumari R, Ekbal A. Amfb: attention based multimodal factorized bilinear pooling for multimodal fake news detection. Expert Systems with Applications, 2021, 184: 115412. https://doi.org/10.1016/j.eswa.2021.115412

24. Song C, Ning N, Zhang Y, et al. A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. Information Processing Management, 2021, 58(1): 102437. https://doi.org/10.1016/j.ijpm.2020.102437

25. Dhawan M, Sharma S, Kadam A, et al. GAME-ON: Graph Attention Network based Multimodal Fusion for Fake News Detection. arXiv preprint arXiv:2202.12478, 2022.

26. Wu Y, Zhan P, Zhang Y, et al. Multimodal fusion with co-attention networks for fake news detection. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021: 2560–2569.

27. Xu P, Zhu X, Clifton D A. Multimodal Learning with Transformers: A Survey. arXiv preprint arXiv:2206.06488, 2022.

28. Gavrilyuk K, Sanford R, Javan M, et al. Actor-transformers for group activity recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 839-848.
29. Xu P, Zhu X. Deepchange: A large long-term person re-identification benchmark with clothes change. arXiv e-prints, 2021: arXiv: 2105.14685.

30. Sun C, Myers A, Vondrick C, et al. Videobert: A joint model for video and language representation learning. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 7464-7473.

31. Guo D, Ren S, Lu S, et al. Graphcodebert: Pre-training code representations with data flow. arXiv preprint arXiv:2009.08366, 2020.

32. Shi B, Hsu W N, Lakhota K, et al. Learning audio-visual speech representation by masked multimodal cluster prediction. arXiv preprint arXiv:2201.02184, 2022.

33. Zheng R, Chen J, Ma M, et al. Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation. International Conference on Machine Learning. PMLR, 2021: 12736-12746.

34. Li R, Yang S, Ross D A, et al. Ai choreographer: Music conditioned 3d dance generation with aist++. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 13401-13412.

35. Lin J, Yang A, Zhang Y, et al. Interbert: Vision-and-language interaction for multi-modal pretraining. arXiv preprint arXiv:2003.13198, 2020.

36. Murahari V, Batra D, Parikh D, et al. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. European Conference on Computer Vision. Springer, Cham, 2020: 336-352.

37. Lu J, Batra D, Parikh D, et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems, 2019, 32.

38. Zhan X, Wu Y, Dong X, et al. Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 11782-11791.

39. Tsai Y H H, Bai S, Liang P P, et al. Multimodal transformer for unaligned multimodal language sequences. Proceedings of the conference. Association for Computational Linguistics. Meeting. NIH Public Access, 2019, 2019: 6558.

40. Wang G, Tan L, Shang Z, et al. Multimodal Dual Emotion with Fusion of Visual Sentiment for Rumor Detection. arXiv preprint arXiv:2204.11515, 2022.

41. Azri A, Favre C, Harbi N, et al. Calling to CNN-LSTM for Rumor Detection: A Deep Multi-channel Model for Message Veracity Classification in Microblogs. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham, 2021: 497-513.

42. Kirchknopf A, Sljipecevic D, Zeppelzauer M. Multimodal detection of information disorder from social media. arXiv preprint arXiv:2105.15165, 2021.

43. Wang Y, Ma F, Jin Z, et al. Eann: Event adversarial neural networks for multi-modal fake news detection. Proceedings of the 24th acm sigkdd international conference on knowledge discovery data mining. 2018: 849-857.

44. Khattar D, Goud J S, Gupta M, et al. Mvae: Multimodal variational autoencoder for fake news detection. The world wide web conference. 2019: 2915-2921.