Journal of Intelligent & Fuzzy Systems 45 (2023) 12425–12437
DOI:10.3233/JIFS-233508
IOS Press

Multimodal fusion sensitive information classification based on mixed attention and CLIP model[1]

Shuaina Huang[a,b], Zhiyong Zhang[a,b,*], Bin Song[a,b] and Yueheng Mao[a,b]
[a]*College of Information Engineering, Henan University of Science and Technology, Henan Luoyang, China*
[b]*Henan International Joint Laboratory of Cyberspace Security Applications, Henan University of Science and Technology, Henan Luoyang, China*

Abstract. Social network attackers leverage images and text to disseminate sensitive information associated with pornography, politics, and terrorism,causing adverse effects on society.The current sensitive information classification model does not focus on feature fusion between images and text, greatly reducing recognition accuracy.To address this problem, we propose an attentive cross-modal fusion model (ACMF), which utilizes mixed attention mechanism and the Contrastive Language-Image Pre-training model.Specifically, we employ a deep neural network with a mixed attention mechanism as a visual feature extractor. This allows us to progressively extract features at different levels. We combine these visual features with those obtained from a text feature extractor and incorporate image-text frequency domain information at various levels to enable fine-grained modeling. Additionally, we introduce a cyclic attention mechanism and integrate the Contrastive Language-Image Pre-training model to establish stronger connections between modalities, thereby enhancing classification performance.Experimental evaluations conducted on sensitive information datasets collected demonstrate the superiority of our method over other baseline models. The model achieves an accuracy rate of 91.4% and an F1-score of 0.9145. These results validate the effectiveness of the mixed attention mechanism in enhancing the utilization of important features. Furthermore, the effective fusion of text and image features significantly improves the classification ability of the deep neural network.

Keywords: Multi-modal, sensitive information, spatial attention mechanism, channel attention mechanism, deep learning

1. Introduction

The rapid development of the internet has sparked urgent attention to the security of social network content [1]. The content and dissemination forms of sensitive information related to pornography, politics, and terrorism have become more diverse.Compared to single modality, multimodal sensitive information that combines text and image is easier to spread and more destructive to public safety.Therefore, research on multimodal sensitive information classification [2] can not only help governments quickly identify sensitive information, but also effectively purify the network ecological environment.

Computer vision and Natural language processing are two of the most useful and powerful technologies to deal with the classification of sensitive information.Alshalan et al. [3] proposed a model based on

[1]The work was supported by National Natural Science Foundation of China Grant No. 61972133, Project of Leading Talents in Science and Technology Innovation in Henan Province Grant No. 204200510021, Program for Henan Province Key Science and Technology under Grant No. 222102210177 and Henan Province University Key Scientific Research Project under Grant No. 23A520008.

*Corresponding author. Zhiyong Zhang, Henan International Joint Laboratory of Cyberspace Security Applications, Henan University of Science and Technology, Henan Luoyang, 471023, China. E-mail: zhangzy@haust.edu.cn.

CNN and recursive neural network to detect hate speech on Twitter. Gangwar et al. [4] proposed a deep CNN architecture (AttM-CNN) combining attention mechanism and metric learning for pornographic image classification.The metaherstics also shows significant potential in tackling machine learning hyperparameters optimisation challenge [5]. Zare et al. [6] proposed a method based on Diferential Evolution (DE)/currentto-best/1 for enhancing the FA's movement process. The proposed modifcation increases the global search ability and the convergence rates. Multimodal classification [7] which aims to classify a pair, or a group, of different types of data into categories, has attracted extensive attention from academia and society.However,the sensitive information in the form of combining images with text limits the accuracy of single mode recognition.Current multimodal methods for capturing the components of multimodal contexts are too simple to extract high-order complementary information from multimodal contexts.

In order to address these challenges, our study proposes an innovative multimodal fusion model called ACMF . ACMF incorporates a hybrid attention mechanism and a transformer architecture to effectively capture the correlation between image and text, as well as the significance of abstract image features across various regions and levels. The primary contributions of our study are summarized as follows:

(1) the image features were extracted utilizing the hybrid attention mechanism, which effectively mines relevant information. On the other hand, the text features were modeled using the state-of-the-art transformer structure, which has demonstrated remarkable performance in various natural language processing tasks. To fuse the image and text features, we mapped the image feature representation domain to align with the text feature representation domain. Additionally, we employed a multimodal adaptive analysis method to capture and model the correlation between image and text information.
(2) By incorporating two attention mechanisms into our feature extraction process, we aim to improve the representation and discrimination capabilities of the image features, ultimately enhancing the overall classification performance of the sensitive information detection model.
(3) A feature fusion module based on cyclic attention mechanism is constructed, which can effectively promote the interaction between image features and text features. Moreover, the fusion effect of features is further improved through multiple cycle fusion. The multi-task joint learning combined with CLIP task not only ensures the full utilization of graphic features, but also enhances the correlation between graphic features, and carries out weighted summation of graphic feature loss and classification loss, so as to enhance the efficiency of online learning.

The rest of this paper is organized as follows: Section2 introduces the overview of related approaches. Section 3 describes the proposed method in detail. Section 4 presents the datasets, experimental setup, baseline and the results obtained from applying the proposed method.Section 5 presents the conclusion of this study .

2. Related work

2.1. Classification of single-modal sensitive information

Most studies on sensitive information classification have focused on single- modal data, such as image classification or text classification. CNNs can automatically extract abstract features from the original pixels [8] and identify image information [4] and have thus made remarkable achievements in the field of image processing. In addition, various swarm intelligence algorithms improve the global search ability and the convergence rates, which is very enlightening to the algorithm optimization of artificial intelligence [9, 10]. Banaeeyan et al. [11] used the residual neural network (ResNet) to detect nudity in an automated process. Perez et al. [12] proposed a pornographic video classification method by using CNN and static motion information.

Natural language processing is witnessing rapid advances and yields excellent results in tasks such as semantic analysis , sentiment analysis , and sentence modeling [13]. Feature selection and feature extraction methods are also more optimized [14, 15]. Pan et al. [16] proposed a sentiment analysis model for Chinese online course reviews by using an efficient transformer that enables parallel input of word vectors for predicting sentiment polarization. Zivkovic

et al. [17] proposed an arithmetic optimization algorithm (AOA) - based approach that can improve the classification of fake news results by reducing the number of features and achieve high accuracy. Kar et al. [18] proposed an optimal feature extraction and hybrid diagonal gated recurrent neural network (FE-DGRNN) for hate speech detection and sentiment analysis in multilingual code-mixed texts.Agushaka et al. [19] proposed a novel population-based metaheuristic algorithm called the Gazelle Optimization Algorithm (GOA), inspired by the gazellesâŁ™ survival ability in their predator-dominated environment.The results show that GOA is a potent tool for optimization that can be adapted to solve problems in different optimization domains.Hu et al. [20] proposed an adaptive hybrid dandelion optimizer called DETDO to address the shortcomings of weak DO development, easy to fall into local optimum and slow convergence speed.

Single-modal algorithms have limitations in real-world sensitive information classification because sensitive information in social networks mainly exists in the hybrid form of images and texts. Traditional single-modal algorithms cannot deal with multi-modal information. In response to this, in this study, we designed a more targeted algorithm based on multimodal sensitive information to improve the sensitive information classification performance.

2.2. Classification of multimodal sensitive information

The outstanding performance of deep neural networks (DNNs) in single-modal recognition for sensitive images, texts, audios, and videos enables reidentifying fused multimodal sensitive information. Multimodal feature fusion classification allows the correction of the internal error from a single information source [21] and combines the features extracted from each mode, thus achieving more accurate and robust classification. Weng et al. [22] proposed an XMATL frame- work based on MAT-LSTM that combines text features with user behavior to detect toxic articles (e.g., rumors, pornography, and fraudulent content) on social networks. To improve the effectiveness of multilingual toxic comment detection, Song et al. [23] proposed a hybrid model combining monolingual and multilingual models that can fuse monolingual and multilingual features from different BERT models. Zhang et al. [2] proposed an end-to-end deep learning (DL)-based multimodal fusion network that splices image-text features by using the attention mechanism and has been employed for multimodal rumor detection on social networks.

According to the studies, the multimodal algorithm improves model stability and prevents errors in a single mode from affecting the model judgment. Image data contains visual information while text data contains semantic information. Fusion of these two types of data utilizes their respective advantages and achieves complementary effects. Multimodal information processing facilitates mining of hidden sensitive information and reduction of recognition errors. However, existing methods do not consider the importance of image region features and channel features, resulting in insufficient mining of image and text features and lack of mode correlation establishment. In sensitive information detection, decision-level fusion makes it difficult for the model to capture the correlation between image and text features. Based on this, the ACMF model is proposed. It uses a mixed attention mechanism and Bert to extract image and text features, constructs a cyclic attention feature fusion module after cross combination of text and image features, and compares the similarity of text and image using the CLIP model to optimize the network weight of text and image feature extraction. Finally, multi-modal features are classified after weighted summation to improve the classification accuracy of multi-modal sensitive information.

3. Proposed Architecture: ACMF

The ACMF model, as depicted in Fig. 1, comprises several key components for effective multimodal sensitive information classification. These components include an image feature extraction module, a text feature extraction module, a cycle attention feature fusion module, a CLIP text-image feature similarity comparison module, and a multimodal sensitive information classification module.To begin with, image features and text features are extracted using a convolutional neural network (CNN) and BERT (Bidirectional Encoder Representations from Transformers) respectively. These extraction modules incorporate a mixed attention mechanism to capture relevant information from both modalities.Next, the cross-fusion of image features and text features is performed, resulting in hybrid features that encompass both visual and textual information. This fusion enables a comprehensive representation of the multimodal content. The cyclic attention mech-

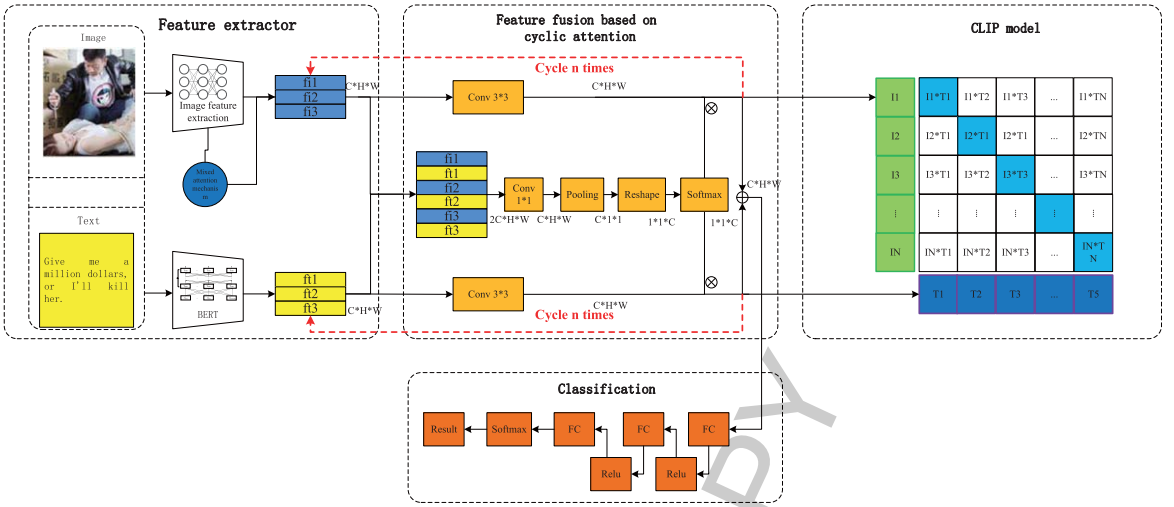


Fig. 1. Structure of the ACMF model.

anism is then employed to facilitate the weighted fusion of image features and text features. This mechanism ensures effective interaction and information transfer between the two modalities, enhancing the overall representation ability of the features.To further enhance the representation and similarity analysis of the multimodal features, the CLIP module is introduced. The CLIP module enables similarity matching between image and text features, strengthening the relationship between the two modalities. Finally, based on the fused multimodal features, the multimodal sensitive information classification module classifies the data into categories such as pornography, politics, terrorism, and others.The comprehensive architecture of the ACMF model, as outlined above, provides a robust framework for capturing and integrating multimodal features, enabling effective classification of sensitive information.

3.1. Image feature extraction

In image processing, traditional CNNs generally accord the same attention to each region in an image. However, in classification tasks, different regions of the image contribute differently to the task. Thus, the focus should be on the regions that are related to the task. For example, in the classification of terrorism-related images, features such as guns and ammunition require special attention; thus, regions related to these features should be assigned a higher weight in discrimination, whereas background regions should be assigned a lower weight. The correlation between feature regions and classification tasks should be learned when extracting image features; thus, in the proposed model, the spatial attention mechanism is used.

The proposed ACMF model fuses the shallow and deep image features through residual connection, thus improving the feature utilization efficiency of the model. However, different channels of the CNN extract different levels of abstract features. Traditional CNNs treat the output feature of each channel equally and cannot process the information of the channel features. In view of this problem, in the proposed model, the channel attention mechanism is adopted and the importance of each channel is determined for the categorization task. This channel attention mechanism is combined with the spatial attention mechanism to form the hybrid attention mechanism model.

3.1.1. Spatial attention mechanism model

Suppose there are I multimodal sensitive information image-text samples $\chi = \{x_1, x_2, \cdots, x_I\}$, For the i -th sample, the classification label is $y_i \in \{y_{i0}, y_{i1}, y_{i2}, y_{i3}\}$, where $y_{i0}, y_{i1}, y_{i2}, y_{i3}$ represent the four categories: normal, terrorism-related, political, and pornographic. Each multimodal sensitive information is divided into two modes $x = \{x^v, x^t\}$, where superscripts v and t represent visual and textual forms, respectively. For a given sample of multimodal sensitive information x_i,the task is to classify it correctly. In CNNs, the feature space of image x_i^v after processing by several hidden layers can be divided into several regions denoted by $R_i^v = \left\{r_{i(1)}^v, r_{i(2)}^v, \cdots, r_{i(D)}^v\right\} \in R^{D\times C}$, where $D = w \cdot h$ is

the number of image feature regions, C is the number of feature region channels. The process can be formulated as follows:

$$R_i^v = f_s(x_i^v; \theta_c),\ R_i^v \in R^{D \times C} \tag{1}$$

where θ_c is the parameter of the convolutional layer, and f_s is convolutional mapping composed of several hidden layers. The contributions of different feature regions $r_{i(j)}^v$ to the classification task are different. The spatial attention mechanism learns the importance of the various feature regions in an image by establishing a neural network model. The feature region F_i^s after the addition of the spatial attention mechanism can be obtained from the weighted sum of the learned feature region importance and the feature region. Compared with the original feature region, F_i^s assigns more weight to the feature regions that have a strong correlation with the classification result; this aids in achieving better classification results. The whole process can be expressed as

$$\begin{aligned} \tilde{R}_i^v &= W_{s3} * \left(R_i^v + W_{s2} * \left(W_{s1} * R_i^v\right)\right) \\ \hat{r}_{i(j)}^v &= \delta\left(\tilde{r}_{i(j)}^v\right) \\ F_i^s &= \sum_j \hat{r}_{i(j)}^v \cdot r_{i(j)}^v \end{aligned} \tag{2}$$

where $R_i^v = \left\{r_{i(1)}^v, r_{i(2)}^v, \cdots, r_{i(D)}^v\right\} \in R^{D \times C}$ is the input image feature, $*$ represents two-dimensional $2d$ convolution, and W_{s1}, W_{s2}, W_{s3} are convolution operations with a kernel size of 3×3. After multiple rounds of convolution of the abstract features, the feature regions are converted to $\widetilde{R}_i^v = \left\{\widetilde{r}_{i(1)}^v, \widetilde{r}_{i(2)}^v, \cdots, \widetilde{r}_{i(D)}^v\right\} \in R^{D \times 1}$, where $D = \omega \cdot h$ is the number of image feature regions, 1 is the number of feature region channels, $\widetilde{r}_{i(j)}^v \in \widetilde{R}_i^v$, and δ is the sigmoid function used for extracting the weights of the various feature regions. Finally, the weighted image feature $F_i^s \in R^{D \times C}$ is obtained as the output. The network structure of the spatial attention mechanism is illustrated in Fig. 2.

3.1.2. Channel attention mechanism model

After several layers of convolution, the output features F_i^s of image x_i^v after the spatial attention mechanism are listed according to the channel sequence as $U_i = \left\{u_{i(1)}, u_{i(2)}, \cdots, u_{i(c)}\right\} \in R^{D \times C}$, where $D = w \cdot h$ and C is the number of characteristic channels. Each layer of the convolutional network includes several kernels that have a local acceptance domain. However, every cell in U_i cannot leverage the contextual information outside that domain. To solve this problem, in the proposed model, convolution is used to compress the global spatial information to generate channel statistics, and the global spatial feature of each channel is taken as its representation:

$$\widetilde{U}_i = W_{c3} * (U_i + W_{c2} * (W_{c1} * U_i)) \tag{3}$$

where $*$ represents $2d$ convolution, W_{c1}, W_{c2} are 3×3 convolution operations, W_{c3} is $w\times$ hconvolution, and $\widetilde{U}_i \in R^{1 \times c}$.

The results obtained from the above operations are used to determine the dependencies between the channels. The dependencies are then combined with the feature maps of various feature channels to obtain channel features with adaptive importance. This process can be expressed as follows:

$$\hat{u}_{i(j)}^v = \delta\left(\tilde{u}_{i(j)}^v\right),\ F_i^v = \sum_j \hat{u}_{i(j)} \cdot u_{i(j)} \tag{4}$$

where $\widetilde{u}_{i(j)} \in \widetilde{U}_i$, δ is the sigmoid function, $\hat{u}_{i(j)} \in \widetilde{U}_i$ is the weight of the j -th channel, and $F_i^v \in R^{D \times C}$ is the output feature of the weighted channel. The process is illustrated in Fig. 3.

3.2. Text feature extraction

The text feature extraction module utilizes a pretrained BERT [24] structure for fine-tuning to extract sensitive text features and employs a 12 two-way transformer layer stack. The transformer structure obtains the encoding of each word in a sentence, which contains the dependency of the word on other words in the sentence, and assigns it a global acceptance domain. In addition, the proposed model permits parallel computation.

The text information in the i -th multimodal sensitive information sample is denoted as $x_i^t = \left\{x_{i(1)}^t, x_{i(2)}^t, \cdots, x_{i(s)}^t\right\}$, where $x_{i(s)}^t$ is the word vector representation of the s-th word in the i-th text sample. The output feature $h_{i(j)}^l$ of each word vector $x_{i(j)}^t$ in the l-th hidden layer is iteratively calculated in the transformer layer. $h_{i(j)}^l$ are then aggregated to form a feature matrix H_i^l. The transformer layer consists of two sublayers: a multi-head self-attention sublayer and a feed-forward neural network.

$$\begin{aligned} head_i &= Attention(H_k^l W^Q, H_k^l W^K, H_k^l W^V) \\ \widetilde{H}_i^l &= [head_1, head_2 \ldots, head_h]\, W^0 \end{aligned} \tag{5}$$

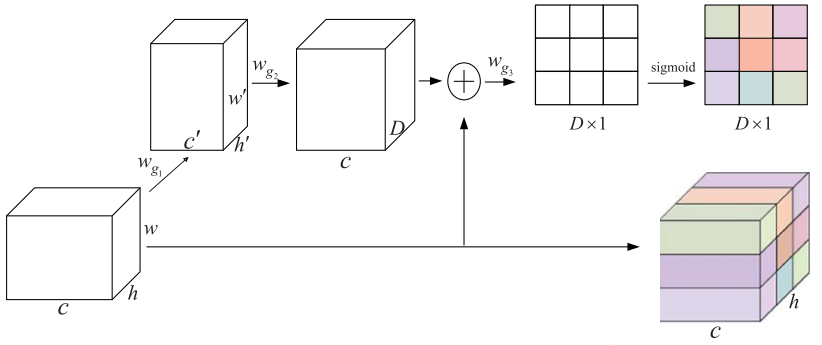


Fig. 2. Structure of spatial attention mechanism.

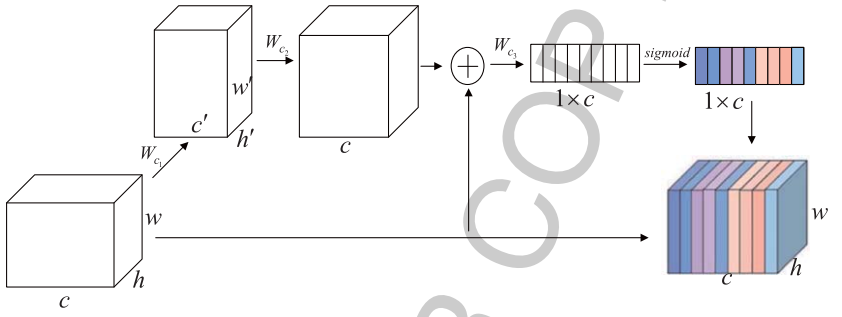


Fig. 3. Structure of channel attention mechanism.

where $Q \in R^{n\times d}$ is the query matrix, $K \in R^{n\times d}$ is the key matrix, $V \in R^{n\times d}$ is the value matrix, and d is the number of cells in the hidden layer of the CNN. The results in $\widetilde{H}_i^l$ are merged, and a feed-forward neural network is applied to the merged output eigenvector matrix. After 12 layers of such transformation, the output feature $F_i^t = \left\{\widetilde{x}_{i(1)}^t, \widetilde{x}_{i(2)}^t, \cdots, \widetilde{x}_{i(s)}^t\right\}$ of all the items in the input text sequence is obtained.

3.3. Image-text multimodal cycle attention feature fusion

In order to improve the interaction of feature information of image text and make full use of pixel features and text features as far as possible, this paper designs a set of circular attention feature fusion module. Now the image features F_i^v are obtained by the mixed attention mechanism, and the text features F_i^t are extracted by Bert structure. At the fusion layer, the two types of feature Spaces are firstly spliced into a multi-modal representation:

$$F_{2i}^m = shuffle(F_i^t; f_v(F_i^v)) \tag{6}$$

where $F_i^v, F_i^t \in R^c$ are the image feature vector representation based on mixed attention mechanism and text feature vector representation based on Transformer, respectively. c is the dimension of the embedded feature vector. $f_v(\cdot)$ is a mapping function approximated by a multilayer neural network. In order to better ensure the interaction and transmission between features $F_i^v and F_i^t$, the weight representation of fusion features is constructed by combining features F_{2i}^m and spatial attention mechanism:

$$F\_att_1^m = softmax(pool(conv_{1*1}(F_{2i}^m))) \tag{7}$$

where $conv_{1*1}$ is $1*1$ convolutional neural network, whose purpose is to carry out channel dimension reduction operation on the original cross-splicing features, *pooling*() is the pooling operation on the feature scale. In order to avoid the problem that the model gradient is difficult to backpropagate due to too large or too small weight features, in this paper, *softmax* activation normalization is carried out on the features after the modified scale.

In order to ensure the high efficiency of multimodal feature fusion, this paper designs a cycle attention structure such as Formula (8) to ensure more adequate transmission and interaction of mixed fea-

tures.

$$f_c = cyclic_n(F\_att_1^m \odot F_i^v, F\_att_1^m \odot F_i^t) \quad (8)$$

where $cyclic_n()$ is the feature fusion operation of cycle attention mechanism, n is the number of cycles and $\odot$ is the dot product of the same position feature and weight.

3.4. Image-text feature match

In order to enhance the image-text feature association, CLIP model is introduced in this paper, and BERT word embedding structure is adopted to map the words or phrases in the vocabulary to the real vector T1-TN. Then the picture is mapped to the feature vector I1-IN through attention mechanism module, so as to construct the image-text feature mixing matrix for fusion. Finally, by comparing the similarity between the text features and the image features, the feature weight with stronger ability to represent the text features is obtained. Its confusion matrix is shown in Fig. 4.

The calculation process of CLIP network is as follows:

$$\begin{aligned} L2\_N &= \min \sum \left(y_i - w^T x_i\right)^2 + \lambda \|w\|_2^2 \\ I_f &= L2\_N(F_i^v \odot w_i^v) \\ T_f &= L2\_N(F_i^t \odot w_i^t) \\ logits &= (If \odot Tf) * et \\ L_{(i,t)} &= L_{ce}(logits, y_{label}) \\ L_c &= \frac{1}{2}(L_i + L_t) \end{aligned} \quad (9)$$

where $L2\_N$ represents $L2$ regularization method; w_i^v and w_i^t respectively represent the weight of learning in the image-text coding network; $logits$ represents the cosine similarity calculation results of the regularized image and text features; L_{ce} represents the cross entropy loss function.

3.5. Multimodal sensitive information classification

The weights of the image and text features learned using the multimodal adaptive neural network are used for obtaining the weighted sum of the image-text features. After outputting the multimodal fusion features, three fully connected layers are added as hidden layers to enable the model to capture the nonlinear high-order correlation between the multimodal

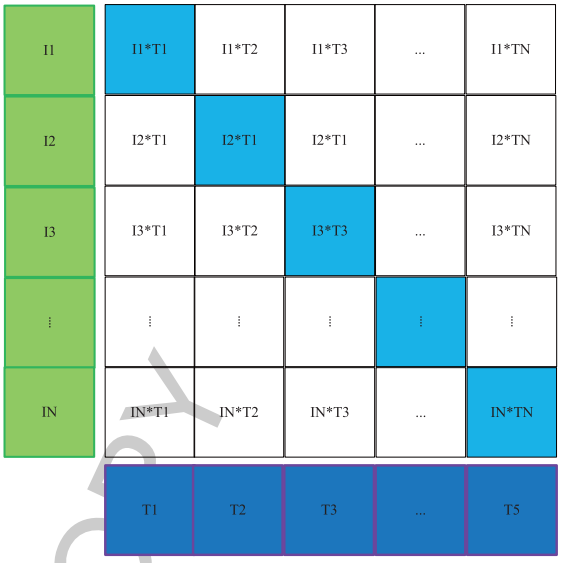


Fig. 4. Text and image feature fusion matrix.

fusion features and sensitive information classes. The formalized definition of the hidden layer is as follows:

$$\begin{aligned} p_0 &= w_0(\partial(m_v) F_i^v + \partial(m_t) F_i^t) + b_0 \\ p_1 &= ReLU(w_1 p_0 + b_1) \\ p_2 &= ReLU(w_2 p_1 + b_2) \end{aligned} \quad (10)$$

where F_i^v, F_i^t are the image and text feature representations, respectively; $\partial(m_v)$, $\partial(m_t)$ correspond to the image and text mode weights, respectively; w_0, w_1, w_2 and b_0, b_1, b_2 are learnable weight matrixes and bias vectors, respectively; and p_0, p_1, p_2 the output neurons of the corresponding hidden layers. $ReLU$ is used as the nonlinear activation function. Finally, the output of the hidden layer p_2 is converted to a classification prediction score by using the following formula:

$$\hat{y}_{ij} = softmax(w_p^T p_2) \quad (11)$$

where w_P is the weight of the prediction layer. The softmax function is used during normalization to obtain the probability distribution predicted by the model. Then, the obtained probability is combined with the cross entropy loss function to obtain the classification result loss of multi-mode mixed features, as shown in the following formula:

$$L_m = \sum y_{label} log(\hat{y}_{ij}) \quad (12)$$

Finally, the loss function of the CLIP task in Section 3.4 is simultaneous to obtain the loss function

of the overall multitask network of ACMF, which is expressed as follows:

$$L = \frac{1}{2}(L_c + L_m) \tag{13}$$

4. Experiments

In this section, we first describe the data sets and experimental setup. Then, we evaluate our proposed method and compare it with other state-of-the-art methods. Finally, we show the ablation study used to investigate our proposed method.

4.1. Datasets

To illustrate our work, we collected a real-world set of data from the Internet, known as the NSASI datasets, which includes 11694 items. Each image-text pair was labeled by three annotators. A majority vote was employed to filter inconsistent entries. The data set contains 3596 pornographic information,2001 political information, 2234 terrorism information and 3863 normal information.The ratio of training set to test set is 7:3. The details of data usage in this paper are as follows: 1) only the image part of NSASI dataset is used for single-modal image sensitive information classification; 2) Single-modal text sensitive information classification only uses the text part of NSASI dataset; 3) Multi-modal image-text sensitive information classification uses the full image-text data of NSASI dataset.

4.2. Experimental settings

The experimental environment uses Windows 10 operating system, NVIDIA GeForce 1080Ti graphics card for model training and testing . CPU configuration is Intel(R) Core(TW) i7-10700CPU, CUDA version 10.1, Python language environment is 3.8 model framework using Pytorch deep learning framework version 1.11.0.

Accuracy, precision, recall, F1- score, and receiver operating characteristic (ROC) curve were used to evaluate the performance of the models.In this paper, the training parameters of ACMF model were determined by combining the previous scholars' parameter tuning experience and a large number of experimental comparisons. The maximum sentence length in the BERT Chinese pretraining language model was set as 32,the batch size was set as 8 during training, the learning rate was 5e-5, a 12-layer transformer structure was selected, the dimension of text features obtained from BERT was 768,the image was rgb color image, the training scale was 224*224*3.In the network, a leaky ReLu activating function is applied to all fully connected layers, while the value of dropout is 0.5 for the purpose of avoiding overfitting. Adam optimizer is also applied to better select the model parameters.

4.3. Baselines

In order to evaluate the performance of the proposed method, the prediction results on the the above data set is quantified and compared with the results of single-modal sensitive information detection models. IM denotes image modality, TM means text modality, and IM-TM corresponds to a multimodal form that combines images with texts.

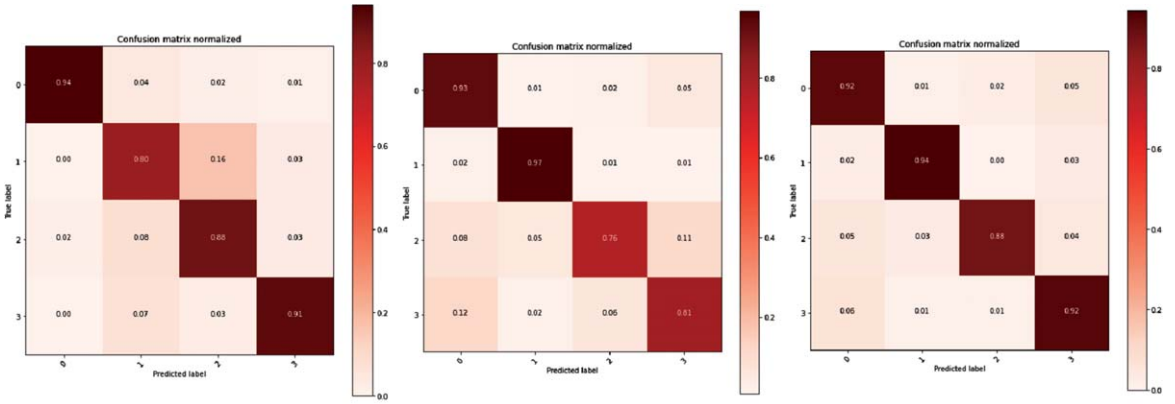
The proposed model was compared with popular image classification benchmark models, namely ChildNet [25], ResNet-34 [26], DenseNet [27] and EfficientNetV2 [28] and text classification benchmark models, namely TextCNN [29], and LSTM [30].

4.4. Performance comparison

As shown in Table 1, ACMF is significantly better than the baseline method in most indicators, indicating that the mixed attention mechanism and cyclic attention mechanism proposed in this paper can effectively improve the performance of sensitive news detection, and the method combining CLIP model with image and text information can improve the detection accuracy. The effect of multimodal fusion model is better than single modal model. When the multimodal sensitive datasets was used, the accuracy of the ACMF model was 12%, 4%, 7.2% and 4.8% higher than that of Child-Net, ResNet-34, TextCNN and LSTM respectively, indicating that multimodal feature fusion can considerably improve the sensitive information classification performance. CLIP model and cyclic attention mechanism can learn the correlation between image and text features, thereby improving the classification accuracy. The single-modal image feature was used for classifying the four types of information, the accuracy of Child-Net, ResNet-34, DenseNet and EfficientNetV2 was between 79.4% and 88.2%, and the performance of EfficientNetV2 was considerably better than that of Child-Net and ResNet-34, thereby reflecting the

Table 1
The performance of each model on sensitive datasets

Method	Modality	Accuracy	Precision	Recall	F1-score
ChildNet [24]	IM	0.7940	0.8107	0.7918	0.7956
ResNet-34 [25]	IM	0.8740	0.8730	0.8724	0.8725
EfficientNetV2 [27]	IM	0.8820	0.8835	0.8811	0.8819
Densenet [26]	IM	0.8720	0.8796	0.8715	0.8712
Text-CNN [28]	TM	0.8420	0.8491	0.8396	0.8407
LSTM [29]	TM	0.8660	0.8682	0.8655	0.8642
ACMF	**IM+TM**	**0.9140**	**0.9162**	**0.9145**	**0.9145**



a) EfficientNetV2 b) LSTM c) ACMF

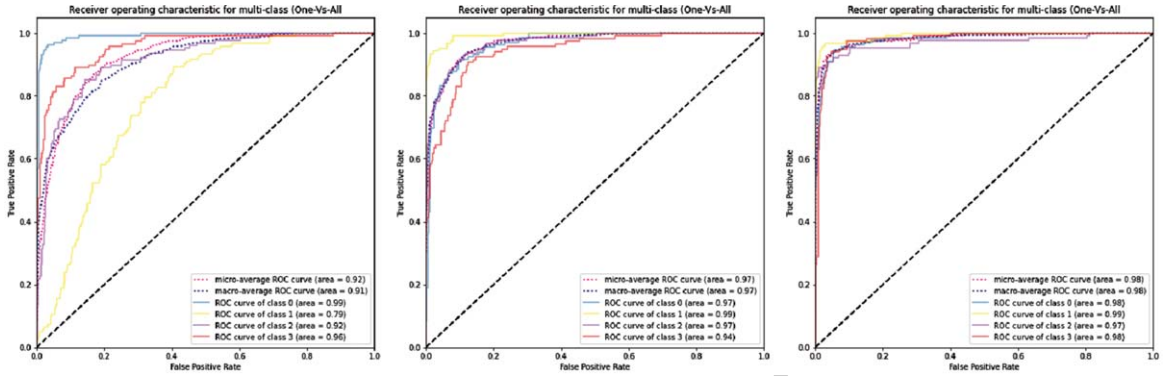
Fig. 5. Confusion matrices for different models.

effectiveness of deep learning methods in sensitive image detection. In the NSASI datasets, the accuracy of EfficientNetV2, ResNet-34 and Densenet was 3% 4% higher than that of TextCN and LSTM,which intuitively illustrates the greater impact of image features on the accuracy of model detection. The accuracy of text detection is low, which may be related to Chinese homophones and polysemy. Therefore, if the detection of sensitive information only relies on text features, it will lead to high false positive rate and degradation of model performance. Combining the rich visual information in the image features with the text features can effectively improve this problem.

To study the correlation between categories, a confusion matrix was introduced to visualize the fine-grained classification performance of models. Each column represents the predicted category, and each row represents the actual category. The confusion matrixes (Fig. 5) of the single-modal models, namely EfficientNetV2 and LSTM, and the multimodal fusion ACMF model were experimentally compared. The ACMF model exhibited a higher classification accuracy than other models in distinguishing sensitive information. Furthermore, the classification effect of ACMF was relatively even and did not exhibit a particularly inferior accuracy for a certain class.

The complexity of the ACMF model was measured using the number of parameters and floating point operations per second (FLOPS). As shown in Table 2, the multimodal ACMF model modified using the BERT model had slightly more parameters than the original model, but it was acceptable. From the perspective of FLOPS, ACMF exhibited great advantages over other common visual environment models, namely ResNet50 and ResNet101, and did not require much in terms of hardware resources. Therefore, the training and prediction of the proposed ACMF model are not time-consuming, and the model can be deployed and run in real-world scenarios.

The model classification performance was also studied using the ROC curve (Fig. 6). The ACMF model had a smoother ROC curve than the other models and exhibited a considerably improved area under the curve (AUC), indicating that the ACMF multimodal fusion classification model had higher robustness and stability in the task of classifying multimodal sensitive information of the four categories.



a) ChildNet b) TextCNN c) ACMF

Fig. 6. ROC curves of different models.

Table 2
Comparison of model complexity

Model	FLOPS	Parameters
ResNet50	3.860G	25.6M
ResNet101	7.341G	44.5M
BERT	2.552G	102.4M
ACMF	**2.695G**	**102.6M**

In order to more intuitively show that the ACMF model has better performance in the classification of sensitive image-text information, we randomly select four types (pornography, politics, terrorism and normal) of pictures to show the results, as shown in Table 3. It can be seen from the results that the ACMF model can accurately recognize multimodal sensitive information. In addition, for the similarity of images and texts, the similarity of images and text features of the same category can reach more than 0.96, which shows that the scheme designed in this paper has good performance.

4.5. Ablation analysis

In order to study the effectiveness of each module proposed in model ACMF, we perform separate ablation analyses by removing some modules. As can be seen from Table 4, experimental results revealed the advantages of the attention mechanism in classification, especially Model-cyl+att with mixed attention and multimodal cycle attention, the accuracy of which was 2.4% and 1.6% higher than that of Model-cyl+CLIP and Model-att+CLIP and the precision was 0.0254 and 0.0182 higher than that of Model-cyl+CLIP and Model-att+CLIP, respectively. The ACMF model that employed image-text classification on the multi modal sensitive datasets demonstrated an accuracy increase of 4.2% and 5.2% compared with ACMF-I and ACMF-T, respectively. These improvements indicate that multimodal features are better than single-modal features for sensitive information classification. The ACMF has improved 1.59 percentage points compared to Model-cyl+mix in F1 score, indicating that the cycle attention feature fusion using adaptive mechanism can improve the model performance. As shown in Fig. 7. The accuracy, precision, recall, and F1-score of ACMF model were better than other models, thus, demonstrating that the multimodal feature fusion classification method with CLIP guided mix attention for sensitive information classification greatly improves the performance of multimodal sensitive information classification. Based on Mixed Attention and CLIP Model greatly improves the performance of multimodal sensitive information classification.

5. Conclusion

This study presents a novel approach for fine-grained multimodal sensitive information classification, aiming to effectively fuse heterogeneous image and text features. The proposed ACMF model combines image feature information and text feature information to enable more accurate identification of multimodal sensitive information.The integration of the mixed attention mechanism and the CLIP model in the ACMF model enables the learning of image and text features with enhanced expression capability. The model adaptively learns the weights of these features, resulting in improved representation of multimodal information. By leveraging these weighted multimodal features, the ACMF model

Table 3
Examples of sensitive analysis of image-text pairs with the proposed module

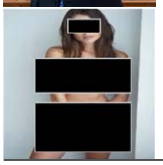
Image	Text	Pornography	Politics	Terrorism	Normal	Image-Text Similarity
	The green grassland makes people feel relaxed.	false	false	false	true	0.9652
	Give me a ransom of one million, or I'll kill him.	false	false	true	false	0.9681
	Obama was elected as President of the United States on November 4, 2008.	false	true	false	false	0.9814
	Sexy mature woman, paid to play any game with	true	false	false	false	0.9767

Table 4
Experimental results of ACMF ablation

Method	Modality	Accuracy	Precision	Recall	F1-score
Model-cyl+CLIP	IM+TM	0.8720	0.8749	0.8723	0.8719
Model-mix+CLIP	IM+TM	0.8800	0.8821	0.8805	0.8802
Model-cyl+mix	IM+TM	0.8960	0.9003	0.8956	0.8964
ACMF-I	IM	0.8720	0.8717	0.8722	0.8719
ACMF-T	TM	0.8620	0.8654	0.8610	0.8619
ACMF	**IM+TM**	**0.9140**	**0.9162**	**0.9145**	**0.9145**

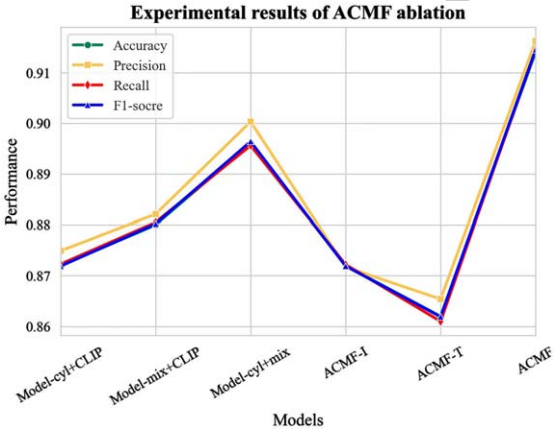


Fig. 7. Experimental results of ACMF ablation.

achieves accurate classification of sensitive information.

The proposed ACMF model introduces a novel approach to multimodal fusion for sensitive information classification, leveraging the mixed attention mechanism and the CLIP model. This model establishes a new baseline in the field, opening up opportunities for further advancements. However, there are some limitations, including dataset, semantic understanding, model scalability, cross-domain application, and security and privacy issues. To overcome these limitations, it is necessary to improve data sets, improve semantic understanding, design more efficient data processing methods, explore cross-domain applications, and strengthen data security and privacy protection. Future studies can focus on fine-tuning the parameters used in the ACMF model to enhance the classification accuracy even further. By optimizing the model's parameters, researchers can potentially achieve improved performance and more accurate identification of sensitive information.

Moreover, with the rapid increase in the number of users sharing short videos on social networking sites, it has become crucial to address the security concerns associated with sensitive content in these videos. Future research can explore the combination of audio, image, and text features to develop effective methods for classifying multimodal sensitive videos. These future research directions have the potential to advance the field of multimodal sensitive information classification and contribute to the development of more comprehensive and robust security systems for social networking platforms.

References

[1] S.S. Roy, A. Roy, P. Samui, M. Gandomi and A.H. Gandomi, Hateful sentiment detection in real-time tweets: An LSTM-based comparative approach, *IEEE Transactions on Computational Social Systems* (2018).

[2] S. Zhang, S. Du, X. Zhang and T. Li, Social rumor detection method based on multimodal fusion, *Computer Sciences* **48** (2021), 117123.

[3] R. Alshalan and H. Al-Khalifa, A deep learning approach for automatic hate speech detection in the saudi twittersphere, *Applied Sciences* **10** (2020), 8614.

[4] A. Gangwar, V. Gonzt'alez-Castro, E. Alegre, et al., Attmcnn: Attention and metric learning based cnn for pornography, age and child sexual abuse (csa) detection in images, *Neurocomputing* **445** (2021), 81104.

[5] M. Zivkovic, M. Tair, K. Venkatachalam, N. Bacanin, Hubálovský and P. Trojovský, Novel hybrid firefly algorithm: An application to enhance XGBoost tuning for intrusion detection classification, *Peer J Computer Science* **8** (2022), e956.

[6] M. Zare, M. Ghasemi, A. Zahedi, K. Golalipour, S.K. Mohammadi, S. Mirjalili and L. Abualigah, A global best-guided firefly algorithm for engineering problems, *Journal of Bionic Engineering* (2023), 1–30.

[7] T. Cheung and K. Lam, Crossmodal bipolar attention for multimodal classification on social media, *Neurocomputing* **514** (2022), 1–12.

[8] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A.L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40** (2017), 834848.

[9] J.O. Agushaka, A.E. Ezugwu and L. Abualigah, Dwarf mongoose optimization algorithm, *Computer Methods in Applied Mechanics and Engineering* **391** (2022), 114570.

[10] A.E. Ezugwu, J.O. Agushaka, L. Abualigah, S. Mirjaili and A.H. Gandomi, Prairie dog optimization algorithm, *Neural Computing and Applications* **34** (2022), 20017–20065.

[11] R. Banaeeyan, H.A. Karim, H. Lye, et al., Automated nudity recognition using very deep residual learning network, *International Journal of Recent Technology and Engineering* **8** (2019), 136141.

[12] M. Perez, S. Avila, D. Moreira, et al., Video pornography detection through deep learning techniques and motion information, *Neurocomputing* **230** (2017), 279293.

[13] N. Kalchbrenner, E. Grefenstette and P. Blunsom, A convolutional neural network for modelling sentences, *52nd Annual Meeting of The Association for Computational Linguistics*, 2014.

[14] M. Stankovic, M. Antonijevic, N. Bacanin, M. Zivkovic, M. Tanaskovic and D. Jovanovic, Feature Selection by Hybrid Artificial Bee Colony Algorithm for Intrusion Detection, *2022 International Conference on Edge Computing and Applications*, 2022, pp. 500–505.

[15] S. Gite, S. Patil, D. Dharrao, M. Yadav, S. Basak, A. Rajendran and K. Kotecha, Textual feature extraction using ant colony optimization for hate speech classification, *Big Data and Cognitive Computing* **7** (2023), 45.

[16] F. Pan, H.B. Zhang, J.C. Dong, et al., Aspect sentiment analysis of chinese online course review based on efficient transformer, *Computer Sciences* **48** (2021), 264269.

[17] M. Zivkovic, C. Stoean, A. Petrovic, N. Bacanin, I. Strumberger and T. Zivkovic, A novel method for COVID-19 pandemic information fake news detection based on the arithmetic optimization algorithm, *2021 23rd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing* (2021), 259–266.

[18] P. Kar and S. Debbarma, Multilingual hate speech detection sentimental analysis on social media platforms using optimal feature extraction and hybrid diagonal gated recurrent neural network, *J Supercomput* (2023).

[19] J.O. Agushaka, A.E. Ezugwu and L. Abualigah, Gazelle optimization algorithm: A novel nature-inspired metaheuristic optimizer, *Neural Computing and Applications* **35** (2023), 4099.

[20] G. Hu, Y. Zheng, L. Abualigah and A.G. Hussien, DETDO: An adaptive hybrid dandelion optimizer for engineering optimization[J], *Advanced Engineering Informatics* **57** (2023), 102004.

[21] L. Lopez-Fuentes, J. van de Weijer, M. Bolanos, et al., Multimodal deep learning approach for flood detection, *MediaEval* **17** (2017), 1315.

[22] Y. Weng, M. Wu, X. Chen, et al., Wechat toxic article detection: A data-driven machine learning approach, *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018.

[23] G. Song and D. Huang, A hybrid model for monolingual and multilingual toxic comment detection, *Tehnicki vjesnik* **28** (2021), 16671673.

[24] J. Devlin, M.W. Chang and K. Lee, Toutanova and Kristina, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018).

[25] R.M. Alguliyev, F.J. Abdullayeva and S.S. Ojagverdiyeva, Image-based malicious internet content filtering method for child protection, *Journal of Information Security and Applications* **65** (2022), 103123.

[26] K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image ecognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[27] Y. Tao, M. Xu, Z. Lu, et al., DenseNet-based depth-width double reinforced deep learning neural network for high-resolution remote sensing image per-pixel classification, *Remote Sensing* **10** (2018), 779.

[28] M. Tan and Q.V. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, (2019).

[29] P. Zhou, Z. Qi, S. Zheng, et al., Text classification improved by integrating bidirectional lstm with two-dimensional max pooling, *26th International Conference on Computational Linguistics*, 2016.

[30] Y. Wang, M. Huang, X. Zhu and L. Zhao, Attention-based LSTM for aspect-level sentiment classification, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 606–615.