Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.056008

ARTICLE

Research on Fine-Grained Recognition Method for Sensitive Information in Social Networks Based on CLIP

Menghan Zhang[1,2], Fangfang Shan[1,2,*], Mengyao Liu[1,2] and Zhenyu Wang[1,2]

[1]School of Computer Science, Zhongyuan University of Technology, Zhengzhou, 450007, China

[2]Henan Key Laboratory of Cyberspace Situation Awareness, Zhengzhou, 450001, China

*Corresponding Author: Fangfang Shan. Email: 6129@zut.edu.cn

Received: 12 July 2024 Accepted: 11 September 2024 Published: 15 October 2024

ABSTRACT

With the emergence and development of social networks, people can stay in touch with friends, family, and colleagues more quickly and conveniently, regardless of their location. This ubiquitous digital internet environment has also led to large-scale disclosure of personal privacy. Due to the complexity and subtlety of sensitive information, traditional sensitive information identification technologies cannot thoroughly address the characteristics of each piece of data, thus weakening the deep connections between text and images. In this context, this paper adopts the CLIP model as a modality discriminator. By using comparative learning between sensitive image descriptions and images, the similarity between the images and the sensitive descriptions is obtained to determine whether the images contain sensitive information. This provides the basis for identifying sensitive information using different modalities. Specifically, if the original data does not contain sensitive information, only single-modality text-sensitive information identification is performed; if the original data contains sensitive information, multi-modality sensitive information identification is conducted. This approach allows for differentiated processing of each piece of data, thereby achieving more accurate sensitive information identification. The aforementioned modality discriminator can address the limitations of existing sensitive information identification technologies, making the identification of sensitive information from the original data more appropriate and precise.

KEYWORDS

Deep learning; social networks; sensitive information recognition; multi-modal fusion

1 Introduction

In the era of digital information, the popularity of social media has brought us countless conveniences and global connections. Through social media, we can stay in touch with family and friends, share life moments, and access various information and entertainment. However, this pervasive digital internet environment also exposes people's privacy to mass disclosure. Just imagine when all your personal information, daily activities, preferences, and even undisclosed inner secrets are constantly monitored and these data are traded among major commercial platforms. Without effective privacy protection, users may face the risk of their personal information being accessed or misused. The content we post on social platforms and our behavioral data become valuable resources for social

media companies to target advertising and provide personalized services. However, the use of this data may also expose our personal privacy and even lead to misuse [1]. Personal information leaks, data security vulnerabilities, and third-party data sharing can all threaten our privacy. For example, in March 2018, The New York Times revealed that Cambridge Analytica collected data from 50 million Facebook users without permission during the 2016 US election and used it to target users with specific political advertisements, influencing the election results [2]. In November 2020, Spotify experienced a data breach where the data of nearly 350,000 accounts were leaked, and all the data was publicly available in a 72 GB database, including over 380 million records [3]. Therefore, in the face of the explosive growth of online data, maintaining a secure environment for the dissemination of information in online communities is a requirement for both domestic and international internet environments.

The key to preventing the leakage of sensitive information lies in accurately and effectively identifying it. Performing single-modal sensitive information recognition on text or images might lead to missing sensitive information, resulting in incomplete recognition. For example, if only text is analyzed for sensitive information, it might not contain any, but the accompanying image could include our personal information. Thus, single-modal processing is partial. On the other hand, directly performing multi-modal fusion on text and images to recognize sensitive information enriches semantics but can also allow irrelevant information to affect the recognition results, leading to inaccuracies. For instance, if a social media post has mismatched text and images, multi-modal sensitive information recognition could lead to errors and increased complexity in recognition. These two methods are commonly used for sensitive information recognition, but both have their shortcomings.

Due to the uncertainties in users' living environments, background experiences, and other factors, sensitive information is manually marked according to different user needs and data attributes to generate a corresponding list of sensitive information. In the process of identifying sensitive information, multi-modal models are not always superior to uni-modal models. Some social dynamics are better suited to multi-modal models, while other information is more suitable for uni-modal models. This paper addresses the aforementioned shortcomings by proposing a modality discriminator. By using the Contrastive Language-Image Pre-training (CLIP), which utilizes its built-in text and image encoders to extract users' social dynamic image information and sensitive descriptions, the modality discriminator performs contrastive learning between images and sensitive descriptions. It identifies sensitive information in images through similarity comparison with sensitive descriptions, achieving sensitivity prediction for images, and screening out images containing sensitive information. The data is divided into images with and without sensitive information. Meanwhile, the modality discriminator converts the sensitivity prediction results of the images into binary classification, exploring the intrinsic connections between different modalities. This determines whether to use uni-modal methods for sensitive information recognition (identifying sensitive information only in text) or multi-modal fusion methods (fusing images containing sensitive information with text in social dynamics for sensitive information recognition).

The main contributions of this paper are as follows:

(1) We propose a modality discriminator for sensitive information recognition based on CLIP. Through text-image contrastive learning, it directly learns from the original user privacy descriptions to calculate the similarity between images and sensitive descriptions.
(2) By utilizing the similarity between images and sensitive descriptions, the classification results are rewritten into the dataset. The dataset is divided into two different modalities for processing: a text dataset and a multi-modal dataset. This approach enables more accurate

sensitive information recognition from the original data and reduces the impact of irrelevant information on sensitive information recognition.

(3) Using two different approaches to handle datasets: for textual datasets, employ single-modal sensitive information recognition, utilizing the BERT model to identify sensitive information in the dataset's text; for multi-modal datasets, employ multi-modal sensitive information recognition, using Vision Transformer (VIT)+Bidirectional Encoder Representations from Transformers (BERT)+Attention to perform multi-modal fusion of images and text in the dataset and identify sensitive information.

2 Related Work

With the emergence and development of social networks, people can now easily and quickly stay in touch with friends, family, and colleagues regardless of their location. Social networks have become a crucial channel for information dissemination, where news, events, and trends can spread globally in an instant, enhancing the efficiency of information dissemination. While users engage with social networking platforms, these platforms also collect personal information, including identity details and social network data. This pervasive digital internet environment has led to significant disclosures of people's privacy. Protecting our social privacy has thus become a major concern in information security.

2.1 Social Network Privacy Protection

In recent years, researchers both domestically and internationally have conducted a series of studies on social network privacy protection. Privacy leaks resulting from information sharing on Online Social Networks (OSNs) are a significant concern for individuals. One of the main culprits behind this issue is the inadequate granularity or flexibility of existing OSN privacy policies, making privacy settings difficult to tailor to individual privacy needs. Yi et al. [4] proposed a new privacy-preserving information-sharing plan for OSNs, where information flow can be controlled according to the privacy requirements of the information owner and the context of the information flow. Specifically, they first formally defined the Privacy Dependency Conditions (PDC) for information sharing in OSNs. Then, based on individuals' heterogeneous privacy needs and potential threats, they designed a PDC-based privacy-preserving information sharing scheme (PDC-InfoSharing) to protect individual privacy. Additionally, to balance information sharing and privacy protection, reinforcement learning techniques were utilized to help individuals achieve this trade-off. Theodorakopoulos et al. [5] proposed a dynamic location histogram privacy approach to focus on the efficiency of different locations being accessed. Ruan et al. [6] proposed an efficient location-sharing protocol that supports location-sharing between friends and strangers while protecting user privacy.

2.2 Text Sensitive Information Identification

Text is the most common form of information sharing in society and the most basic form of data dissemination in social networks. Therefore, identifying sensitive information in text is a fundamental method for protecting user privacy. Social networks have become the most effective platforms for information exchange and are highly favored by internet users. However, the widespread use of social networks also provides a cyberspace for the spread of sensitive content. To address the challenges posed by the large number of deformed and disguised sensitive words in detection, Meng et al. [7] proposed a sensitive word fingerprint convergence method that associates deformed words with the original sensitive words. Finally, for texts containing sensitive words, they developed a convolutional

neural network model based on multi-task learning, which combines sensitivity and sentiment polarity to detect the content of the text. Hassan et al. [8] proposed a more general and flexible solution for protecting text data. They used word embeddings to create vectors that capture the semantic relationships of text terms appearing in a set of documents. Then, they assessed the disclosure caused by text terms on the entity to be protected (e.g., a person's identity or confidential attributes). This method also limited semantic loss (and thus utility loss) by replacing terms in the document with generalized terms for privacy protection, rather than simply suppressing them. Empirical results showed that their approach provided stronger protection and greater utility preservation compared to methods based on named entity recognition, with the additional important advantage of avoiding the burden of manual data labeling. Heni et al. [9] demonstrated a method for identifying sensitive information in Mongo data storage. This method is based on semantic rules to determine the concepts and language components that must be segmented, used to retrieve attributes semantically corresponding to the concepts, and implemented as an expert system for automatically detecting candidate attributes for segmentation. Cong et al. [10] proposed a Chinese sensitive information detection framework based on BERT and knowledge graphs (KGDetector). Specifically, they first trained a pre-trained knowledge graph-based Chinese entity embedding model to describe the entities in the Chinese text input. Finally, they proposed an efficient framework, KGDetector, to detect Chinese sensitive information, which employs a knowledge graph-based embedding model and a CNN classification model. Mehdi et al. [11] proposed a CPSID method to detect personal sensitive information in China, utilizing rule matching to detect specific personal sensitive information consisting only of letters and numbers, and constructing a sequence labeling model named EBC (ELECTRA-BiLSTM-CRF) to detect more complex personal sensitive information composed of Chinese characters. The model employs the latest ELECTRA algorithm to implement word embeddings and uses BiLSTM and CRF models to extract personal sensitive information. By analyzing contextual information, it can accurately detect sensitive entities in China.

2.3 Image Sensitive Information Identification

In the wave of the digital age, an increasing number of images are now shared online on social networking sites such as Facebook, Flickr, and Instagram. Image sharing not only occurs within groups of friends but is also increasingly happening outside users' social circles for the purpose of social discovery. Online image sharing can lead to unnecessary disclosures and privacy infringements. Hou [12] proposed a method for detecting sensitive information in scene images. Given the difficulty of text localization in scene images, this paper specifically proposes an improved SSD algorithm for text localization. Li [13] proposed a sensitive image information detection technology based on comprehensive feature elements, namely the Multi-Information Detection Network (MIDNet), primarily for classifying violent and politically sensitive images. Ashwini et al. [14] proposed a learning model that uses carefully identified image-specific features to automatically predict whether an image's privacy is private or public. They studied deep visual semantic features from various layers of Convolutional Neural Networks (CNNs) and textual features such as user tags and deep labels generated from deep CNNs. They also fine-tuned a pre-trained CNN architecture on their privacy dataset. Shi et al. [15] proposed a homomorphic encryption framework based on efficient integer vectors, applied to deep learning to protect users' privacy in binary convolutional neural network mod, Kaul et al. [16] proposed a knowledge- and learning-based adaptive system for sensitive information identification and processing. Gao et al. [17] utilized image captioning technology to track the propagation of text information in images on the network. Mao et al. [18] proposed a lightweight image-sensitive information detection model based on YOLOv5s. In the feature extraction stage, this

study enhanced the PSA module with an efficient attention module called the GPSA module, enabling the network model to learn richer multiscale feature representations and improving the detection accuracy of sensitive information.

2.4 Multi-Modal Information Recognition

Due to the inadequacy of single-modal text or image-sensitive information detection methods in fully exploring sensitive content in social networks, the trend towards integrating multi-modal methods for sensitive information detection using combined text and image data is inevitable. Hu et al. [19] integrated dynamic fusion modules based on graphs to incorporate multi-modal contextual features from conversations. Yan et al. [20] proposed a multi-tensor fusion network that extracts context-independent uni-modal features, cross-modal feature extraction with multi-modal modeling, and a multi-tensor fusion network for predicting multi-modal emotional intensity. Zou et al. [21] introduced the concept of primary modality and optimized multi-modal fusion through primary modality transformation. Ji et al. [22] proposed an improved Multi-modal Dual-Channel Reasoning mechanism (MDR) that deeply explores semantic information and implicit correlations between modalities based on multi-modal data fusion principles. Additionally, they introduced a Multi-modal Adaptive Spatial Attention mechanism (MAA) to enhance decoder accuracy and flexibility, improving the representation of sensitive information preferences and achieving personalized user privacy preferences. Gao et al. [23], considering global, contextual, and temporal features in videos, captured multiple attentions through a visual attention prediction network. Furthermore, Wang et al. [24] applied multi-modal fusion to similarity-based user recommendation systems, proposing an implicit user preference prediction method with multi-modal feature fusion. Xiao et al. [25] combined visual language fusion with knowledge graph reasoning to further extract useful information. Zhang et al. [26] proposed a new paradigm that does not require sensitive attribute labels, and infers sensitive information by utilizing the visual language model CLIP as a rich source of knowledge, thereby avoiding the need for additional training. Xu et al. [27] argued for considering different modalities for different social media posts in multi-modal information extraction. Multi-modal models are not always superior to unimodal models; some information is better suited to multi-modal patterns, while other information is more suitable for single-modal approaches. Hence, a general data segmentation strategy was proposed to divide social media posts into two sets for better performance under corresponding modal information extraction models.

3 Method

In the process of protecting user social network data, a social network post typically consists of both text and images. In most cases, the text is written by the users themselves. However, concerning the images, the situations described earlier may occur, where the image itself does not contain sensitive information and does not match the text. Alternatively, the combination of the image and text may reveal sensitive information. Given the various types of content shared by users, it is necessary to perform fine-grained partitioning of user data. To achieve this, it is crucial to analyze the sensitive attributes of user images. By categorizing the sensitivity of images, the dataset can be divided into two subsets: a text dataset containing images without sensitive information and a multi-modal dataset containing images with sensitive information. In this process, it is beneficial to employ a Modality Discriminator (CMD) to determine whether the user's image data contains sensitive information. This helps to filter out images that can assist in sensitive information recognition, thereby improving the accuracy of multi-modal sensitive information recognition and reducing the complexity of data processing.

In this research, we have made improvements based on the CLIP model proposed by OpenAI and introduced a modality discriminator. By utilizing the modality discriminator, we process the images in the dataset and divide them into two major parts: the text dataset and the multi-modal dataset. As shown in Fig. 1, different methods are applied to perform sensitive information recognition on these different datasets. For the data in the text dataset, uni-modal sensitive information recognition is suitable. On the other hand, the data in the multi-modal dataset requires multi-modal fusion to accurately identify sensitive information. By combining these two methods, we obtain the final result of sensitive information recognition. The list of sensitive information output is shown in Table 1. The table includes four categories of sensitive information, each containing 10 sensitive topics. When outputting the sensitive list, their labels are displayed.

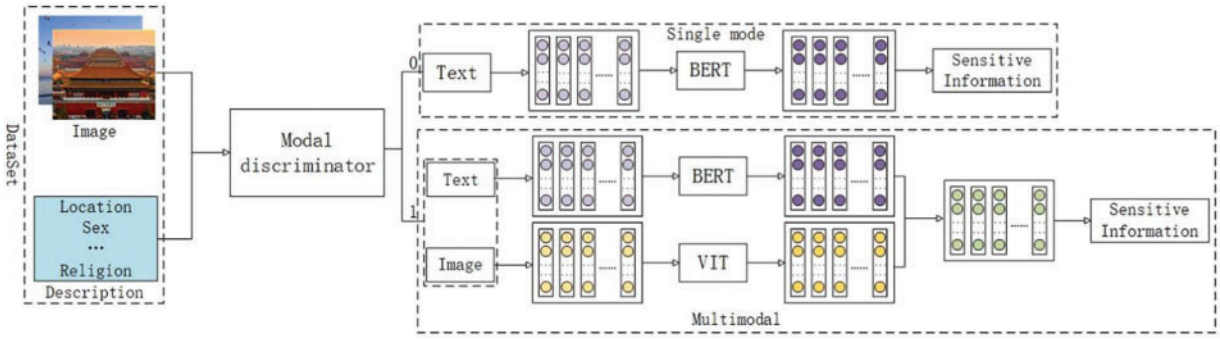

Figure 1: Sensitive information fine-grained recognition framework

Table 1: Sensitive Information List

Label	Location	Sex	Religion	Identity
Related topics	Palestine	Gay	Blasphemy	Wage
	West Bank	Lesbian	Heresy	Pay
	Gaza Strip	Bisexual	Apostasy	Income
	Crimea	Queer	Infidel	Living wage
	Taiwan	Genderqueer	Holy war	Pay secrecy
	Tibet	Non-binary	Religious violence	Doctor
	Xinjiang	Pansexual	Conversion	Lawyer
	South Korea	Asexual	Dlemocrat	Teacher
	North Korea	Biphobia	Sect	Administration
	East Jerusalem	Gay rights	Fundamentalism	Civil servant

3.1 Modal Discriminator

Existing sensitive information recognition techniques typically process data using a single modality, either uni-modal or multi-modal, without simultaneously integrating both approaches. Within these datasets, some information may not contain sensitive content. The presence of such complex and diverse information can make information processing more complicated and can also affect the

final accuracy of the recognition. This section primarily introduces the CMD that performs fine-grained filtering on image information. The dataset is divided into two parts: data suitable for uni-modal processing and data suitable for multi-modal processing. Different recognition methods are then applied based on the results of this filtering. As shown in Fig. 2, we utilize the CLIP model to perform contrastive learning between images and sensitive image descriptions, aiming to uncover the deep connections between images and text. This process helps to identify whether sensitive information exists in the images and outputs corresponding sensitive information for each image. Based on the outputted sensitive information, the dataset is classified accordingly.

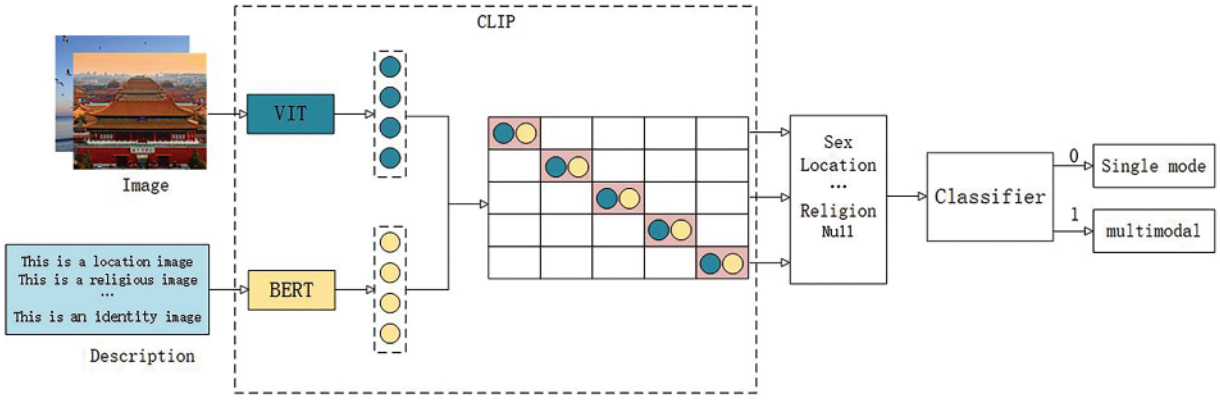

Figure 2: Modal discriminator

Input the sensitive description and image into the input layer of the CLIP model to obtain their semantic vector representations through the encoding layer. The model utilizes a bidirectional encoder architecture, with one Transformer encoder dedicated to processing text data and another VIT encoder responsible for image data. These two encoders share parameters, and the model incorporates a shared embedding space for both text and images. Images and text are mapped into this space, and through contrastive learning of the semantic relationship between images and text, we obtain the embedding vectors for the sensitive description T_i and the image V_i, as shown in Eqs. (1) and (2).

$$T_i = f_{text}(t_i) \tag{1}$$

$$V_i = f_{image}(I_i) \tag{2}$$

where f_{text} is the function for extracting text embedding vectors, and f_{image} is the function for extracting image embedding vectors.

The embedding vectors T_i and V_i are obtained, and then subjected to contrastive learning. The model is trained by maximizing the similarity of positive samples and minimizing the similarity between negative samples. The similarity between the two is calculated using cosine similarity $S(V_j, T_i)$, as shown in Eq. (3).

$$S(V_j, T_i) = \frac{V_j \cdot T_i}{\|V_j\| \, \|T_i\|} \tag{3}$$

The classifier normalizes the similarity using a function to convert it into a class probability. This probability, denoted as $P(y|I_i, T_j)$, represents the likelihood of image I_i belonging to class y,

as shown in Eq. (4).

$$P\left(y = V_j | T_i\right) = \frac{\exp\left(S\left(V_j, T_i\right)\right)}{\sum_{j=1,i=1}^{n} \exp\left(S\left(V_j, T_i\right)\right)} \tag{4}$$

Adding a fully connected layer on top of the aforementioned classifier, the classification is modified to categorize images into two classes: one class indicating sensitive information and the other class indicating non-sensitive information. The output categories are represented as 1 and 0, respectively.

The output of this binary classifier is $\overline{y_{sen}}$, indicating that, given an image and text, it belongs to the sensitive class of images. Eq. (5) is as follows:

$$\overline{y_{sen}} = Sigmoid\left(W \cdot \left(V_j, T_i\right) + b\right) \tag{5}$$

The loss function of this binary classifier is represented as Eq. (6) below:

$$loss = -\left(\overline{y_{sen}} \log\left(\overline{y_{sen}}\right) + \left(1 - \overline{y_{sen}}\right) \log\left(1 - \overline{y_{sen}}\right)\right) \tag{6}$$

where $y_{sen} = 1$ represents the positive class, $y_{sen} = 0$ represents the negative class, 1 indicates that the image contains sensitive information, and 0 indicates that the image does not contain sensitive information.

The cross-entropy loss function is used to compute the contrastive loss, which measures the difference between the predicted similarity distribution and the actual labels' similarity.

3.2 Extracting Sensitive Information from a Single Modality

By using the modality discriminator, the dataset is classified, and the data in the images that do not contain sensitive information undergo text-based sensitive information extraction. The text data in the dataset is preprocessed into word embedding vectors, which are then fed into the BERT model for text classification to output the sensitive information. The process is illustrated in Fig. 3.

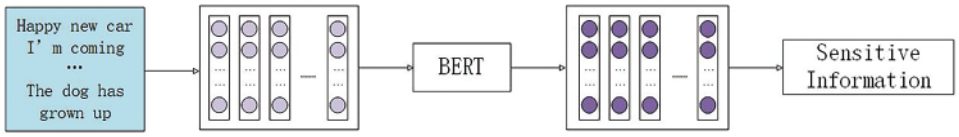

Figure 3: Uni-modal network

The text is tokenized using the tokenizer in BERT, and a special CLS token is added for sensitive classification of the sentence. Each word is converted into its corresponding word embedding. For each text sample i, its word embedding sequence $\left\{W_1^i, W_2^i, \ldots, W_n^i\right\}$ is inputted into the BERT model. The input is processed by the Transformer encoder layer, which utilizes self-attention mechanism to consider the entire input sequence when processing each word. This enables the model to capture dependencies between words and obtain the output vector X_j^i. Eq. (7) is as follows:

$$X_j^i = BERT\left(W_j^i\right) \tag{7}$$

After being processed by multiple layers of encoders, the first vector of the output sequence (i.e., the [*CLS*] token vector) is extracted for the subsequent sensitive classification task. A fully connected layer is added on top of the output layer of the Bert model as a classifier. The [*CLS*] token vector is fed into the classifier, and the softmax activation function is applied to obtain the probability for each

category, as shown in Eq. (8).

$$P\left(y_i = n \middle| \left\{W_1^i, W_2^i, \ldots, W_n^i\right\}\right) = soft\max\left(W \cdot \left[CLS^{(i)}\right] + b\right) \tag{8}$$

where $[CLS^{(i)}]$ represents the hidden state of the special token $[CLS]$ in the output of the BERT model. Its probability distribution represents the probability of the i-th text sequence belonging to category n. y_i represents the true label of the i-th sample.

By calculating the cross-entropy loss function, the difference between the predictions and the true labels is measured to predict the corresponding label for the text, thereby obtaining the sensitive information in the text.

3.3 Extracting Sensitive Information through Multi-Modal Fusion

By using the modality discriminator, a dataset containing images with sensitive information is obtained. The sensitive information extraction is performed through multi-modal fusion on this dataset. The text and images are processed separately using the BERT and VIT models, respectively. The features from both modalities are fused using attention mechanisms, and sensitive information is extracted from the fused features. This process is illustrated in Fig. 4.

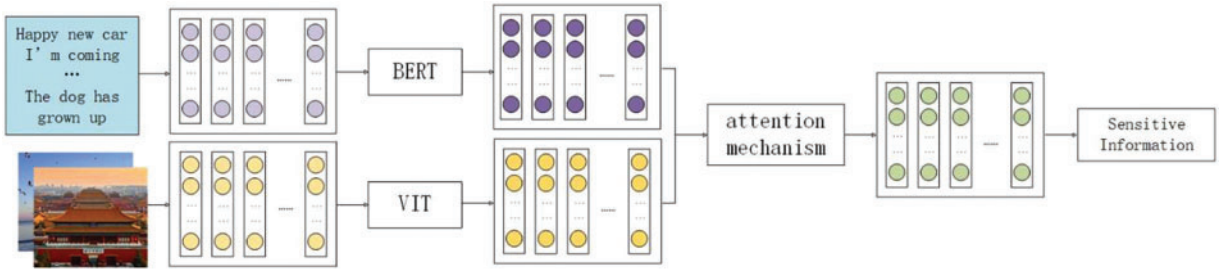

Figure 4: Multi-modal network

The VIT model is based on the Transformer architecture for image processing. It treats an image as a sequence of image patches and processes them to achieve image understanding.

We preprocess our input image by uniformly cropping image $I \in R^{H\times W\times C}(224 \times 224 \times 3)$ to 224 $\times$ 224 pixels and dividing it into image patches of size $P = 32$ pixels. Each patch has a size of 32 $\times$ 32 pixels. The number of patches that can be obtained from the image can be calculated as follows:

$$N = \frac{H \times W}{P^2} \tag{9}$$

After flattening each patch $I_i \in R^{P\times P\times C}$, the resulting vectors are represented as $x_i \in R^{P^2\cdot C}$. The collection of all patches is combined into a sequence, which serves as the input for the VIT model.

$$X = [x_1, x_2, \ldots, x_N] \tag{10}$$

Each flattened patch x_i is mapped to a fixed dimension D through a linear transformation using a learnable embedding matrix E. This linear transformation is achieved by applying the Eq. (11) as follows:

$$Z_i = Ex_i \tag{11}$$

where $E \in R^{D\times(P^2\cdot C)}$ is the weight matrix, and $Z_i \in R^D$ represents the embedding vectors I_{img} after the linear transformation.

Finally, position encoding e_{pos} is added to each patch embedding vector. The representation of the patch vectors is as follows:

$$Z_i = Ex_i + e^i_{pos} \tag{12}$$

where $e^i_{pos} \in R^D$ represents the position encoding for the i-th patch.

The classification token Z_{class} is added at the beginning of the patch sequence to obtain the image vector, as shown in Eq. (13).

$$I_{\text{img}} = ViT\,(I) = [Z_{class}; Z_1; Z_2; \ldots; Z_N] \tag{13}$$

The text is subjected to feature extraction to obtain its text vector, as shown in Eq. (14).

$$T_{\text{txt}} = BERT\,(T) = [h_{class}; h_1; h_2; \ldots; h_M] \tag{14}$$

The text features and image features are transformed through a fully connected layer, as shown in Eqs. (15) and (16).

$$H_{txt} = \text{Re}LU\,(W_{txt}T_{txt} + b_{txt}) \tag{15}$$

$$H_{\text{img}} = \text{Re}LU\left(W_{img}I_{img} + b_{img}\right) \tag{16}$$

where W and b are the parameters of the fully connected layer.

The features are concatenated and fused using attention mechanism to obtain the fused feature H_{fusion}, as shown in Eq. (17).

$$H_{fusion} = TransformerEncoderLayer\left(cat\left(\lfloor H_{txt}, H_{img}\rfloor, \dim = 2\right)\right) \tag{17}$$

The fused feature is processed through a fully connected layer and an activation function, as shown in Eqs. (18)–(20).

$$h_1 = \text{Re}LU\left(W_1H_{fusion} + b_1\right) \tag{18}$$

$$h_2 = \text{Re}LU\,(W_2h_1 + b_2) \tag{19}$$

$$\log its = W_{out}h_2 + b \tag{20}$$

A classification layer is added on top of the fully connected layer. The classification layer maps the fused features to the sensitive information category space, predicts the sensitive information, and calculates the loss using a loss function, as shown in Eqs. (21) and (22).

$$\text{pred\_}label = \arg\max\,(\log its, \dim = 1) \tag{21}$$

$$\text{loss} = CrossEntropyLoss\,(\log its, labels) \tag{22}$$

4 Experiments

4.1 Problem Description

The image below shows two social media posts from a user, which are considered to potentially contain risks. As shown in Fig. 5, the word ‘China’ in the text represents location information, and the overall meaning of the sentence indicates the user’s current state. However, the accompanying

image is unrelated to the text. If we perform multi-modal sensitive information recognition, the image becomes irrelevant and adds complexity to the sensitive information identification. Therefore, this data is better suited for uni-modal sensitive information recognition. By using uni-modal sensitive information recognition, we can extract sensitive information more quickly and accurately.

Figure 5: Problem Description1

As shown in Fig. 6, there is no explicit information in the text pointing to sensitive content. However, the image reveals location information. If we only perform sensitive information recognition on the text, we won't identify any sensitive information since it does not contain any. However, the overall data still compromises the user's privacy by exposing the location information in the image. Therefore, when dealing with similar data for sensitive information recognition, relying solely on uni-modal sensitive information recognition is insufficient to accurately extract sensitive information. In this case, it is necessary to perform multi-modal sensitive information recognition by considering both the text and the image to effectively and accurately detect sensitive information.

To effectively and quickly identify sensitive information within different types of data, we need a more accurate modality selection. Therefore, we propose a modality discriminator that determines the presence of targeted sensitive information in images by comparing the similarity between the images and sensitive descriptions. This approach enables the classification of data instances in the dataset based on their modality. As described in the previous problem statement, data instances without sensitive information are better suited for single-modal processing. On the other hand, data instances containing sensitive information require the fusion of image and text modalities to accurately identify the sensitive information, thus utilizing multi-modal fusion.

From the above problem description, we can observe that there is diversity in the data when identifying sensitive information. To address this diversity, we propose a modality discriminator that performs personalized processing for each data sample. The objective is to achieve more accurate and faster sensitive information recognition.

Figure 6: Problem Description2

4.2 Dataset

Due to the privacy concerns in social network dynamics, there is no publicly available dataset. Therefore, we had to manually collect data for evaluation, using annotated data from 50 students who manually labeled posts on social platforms. Each student annotated 120 pieces of data, including content data ID, text, image, sensitive image description, and sensitive category. This resulted in a total of 6000 pieces of data, with 24,000 individual text data entries. We categorized the collected dataset into sensitive types, divided into four main categories as shown in Table 1 mentioned above, which include location, gender, religion, and identity as sensitive classes. Each sensitive category comprises 10 different topics. For each category, there are 1000 image data samples and 5000 text data samples, with the remaining 2000 image data and 4000 text data samples not containing sensitive information. All the collected information mentioned above was used for training purposes.

4800 image data samples are fed into the CMD, short for CLIP-Modal discriminator, for training. The CMD is then tested using 2000 image data samples. The test dataset is classified using the Modal Discriminator, and the test data predicted as sensitive by the CMD are written into the multi-modal dataset. The remaining test data is written into the text dataset, including content ID and sensitive category. These datasets will be used for subsequent recognition of sensitive information using both single-modal and multi-modal fusion techniques.

4.3 Image Processing

The Bert model is trained using 20,000 text data samples. It is then tested using a general dataset. The multi-modal model (VIT+BERT+Attention), abbreviated as VBA, is trained using 4800 text-image data samples. It is tested using a sensitive dataset.

In traditional convolutional neural networks, the local information of an image is processed through convolutional operations. However, the VIT model divides the image into multiple small patches, treating each patch as a whole, and then feeds these patches into the Transformer for processing. This approach allows for a better capture of the global information in the image, providing a stronger global perception ability to understand the overall structure and content of the image. The image is cropped to 224 × 224 pixels for easy input into the VIT model. Then, the image is converted into a tensor and normalized using the VIT feature extractor for subsequent feature processing, as shown in Fig. 7.

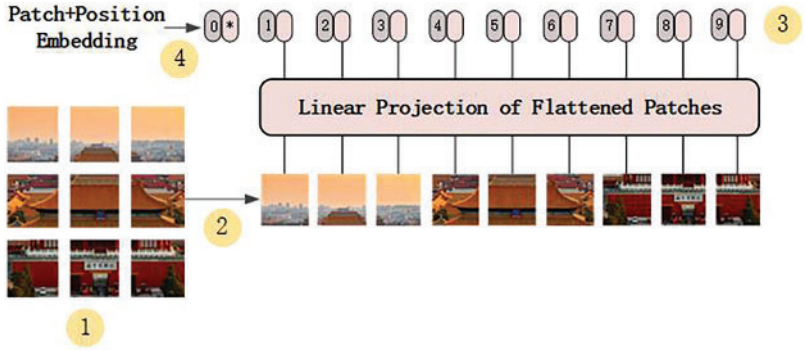

Figure 7: VIT image processing framework

4.4 Results and Analysis

Our proposed model consists of multiple modules, including the Modal Discriminator (CMD), Text Sensitive Information Recognition (T), and Multi-modal Sensitive Information Recognition (VBA). In this experiment, we compared our model with single-modal and multi-modal approaches and evaluated the impact of the Modal Discriminator and multi-modal fusion on the accuracy of the experimental results, as shown in Table 2.

Table 2: The comparison of experimental results

Method	Accuracy	Precision	Recall	F_1
T	0.693	0.750	0.840	0.790
V	0.699	**0.830**	0.730	0.780
T+CMD	0.670	0.750	0.790	0.770
T+CMD+VBA	**0.743**	0.810	**0.850**	**0.830**

Single-Text Model (T): Sensitivity information recognition using only the Bert model for text.

Single-Image Model (V): Sensitivity information recognition using the VIT model for images.

Modal Discriminator Text Recognition (T+CMD): Classifying the dataset using the Modal Discriminator and performing sensitivity information recognition only on the text data within the dataset.

Fine-Grained Sensitivity Information Recognition (T+CMD+VBA): The dataset is divided into text data and multi-modal data using the Modal Discriminator. Sensitivity information recognition is

performed on the text data using the Bert model, while the VBA model is used for multi-modal fusion sensitivity information recognition on the multi-modal data.

The experiments indicate that single-modal recognition of sensitive information in social networks has limitations. However, achieving fine-grained recognition of sensitive information in social networks through the use of a modal discriminator demonstrates better performance in both single-modal and multi-modal scenarios.

5 Conclusion

This paper proposes a modal discriminator that utilizes the CLIP model for contrastive learning between images and sensitive descriptions. It explores the deep semantic connection between text and images and determines the sensitivity of images based on their similarity. This approach enables fine-grained recognition of sensitive information in social networks, reducing the limitations of single-modal sensitive information recognition. It also optimizes the recognition of sensitive information in multi-modal fusion by minimizing the interference of irrelevant information. Moreover, it allows for flexible selection of the optimal processing approach when dealing with different types of data. In the future, this work will be combined with social network access control to eliminate identified privacy issues or establish corresponding access permissions.

In addition to the privacy semantic loss caused by data diversity, another key challenge in protecting online social network data privacy is the dynamic nature of the data. Due to the constant changes in data, it is difficult to ensure privacy protection. As the data for this research has not been publicly available, it hinders the full utilization of the model's advantages in training. In the upcoming research, we aim to address the potential exposure of sensitive information by users in social interactions through visual dialogues.

Acknowledgement: The authors would like to thank the anonymous referees for their valuable comments and helpful suggestions.

Funding Statement: This paper is supported by the National Natural Science Foundation of China (No. 62302540), with author Fangfang Shan for more information, please visit their website at https://www.nsfc.gov.cn/ (accessed on 05 June 2024). Additionally, it is also funded by the Open Foundation of Henan Key Laboratory of Cyberspace Situation Awareness (No. HNTS2022020), where Fangfang Shan is an author. Further details can be found at http://xt.hnkjt.gov.cn/data/pingtai/ (accessed on 05 June 2024). The research is also supported by the Natural Science Foundation of Henan Province Youth Science Fund Project (No. 232300420422), and for more information, you can visit https://kjt.henan.gov.cn (accessed on 05 June 2024).

Author Contributions: Research innovation: Fangfang Shan, Menghan Zhang; Data collection: Mengyao Liu; Data organization and analysis: Zhenyu Wang; Analysis of experimental results: Menghan Zhang; Manuscript writing: Menghan Zhang; Manuscript guidance and revision: Fangfang Shan. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The datasets are available from the corresponding author on reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] G. Akkuzu, B. Aziz, and M. Adda, "Towards consensus-based group decision making for co-owned data sharing in online social networks," *IEEE Access*, vol. 8, pp. 91311–91325, 2020. doi: 10.1109/ACCESS.2020.2994408.

[2] M. Zahak, M. Alizadeh, and M. Abbaspour, "Collaborative privacy management in P2P online social networks," in *2015 12th Int. Iranian Society Cryptol. Conf. Inform. Secur. Cryptol. (ISCISC)*, Rasht, Iran, 2015, pp. 64–72.

[3] E. Palomar, L. González-Manzano, A. Alcaide, and Á. Galán, "Implementing a privacy-enhanced attribute-based credential system for online social networks with co-ownership management," *IET Inf. Secur.*, vol. 10, no. 2, pp. 60–68, Mar. 2016. doi: 10.1049/iet-ifs.2014.0466.

[4] Y. Yi, N. Zhu, J. He, A. D. Jurcut, X. Ma and Y. Luo, "A privacy-dependent condition-based privacy-preserving information sharing scheme in online social networks," *Comput. Commun.*, vol. 200, pp. 149–160, Feb. 2023. doi: 10.1016/j.comcom.2023.01.010.

[5] G. Theodorakopoulos, E. Panaousis, K. Liang, and G. Loukas, "On-the-fly privacy for location histograms," *IEEE Trans. Dependable Secur. Comput.*, vol. 19, no. 1, pp. 566–578, Jan. 1–Feb. 2022. doi: 10.1109/TDSC.2020.2980270.

[6] O. Ruan, L. Zhang, and Y. Zhang, "Location-sharing protocol for privacy protection in mobile online social networks," *EURASIP J. Wirel. Commun. Netw.*, vol. 2021, no. 1, 2021, Art. no. 127. doi: 10.1186/s13638-021-01999-z.

[7] X. Meng and Y. Xu, "Research on sensitive content detection in social networks," *CCF Trans. Netw.*, vol. 2, pp. 126–135, 2019. doi: 10.1007/s42045-019-00021-x.

[8] F. Hassan, D. Sánchez, and J. Domingo-Ferrer, "Utility-preserving privacy protection of textual documents via word embeddings," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 1058–1071, Jan. 1, 2023.

[9] H. Heni and F. Gargouri, "Towards an automatic detection of sensitive information in Mongo database intelligent systems design and applications," in *Intell. Syst. Design Appl.: 18th Int. Conf. Intell. Syst. Design Appl. (ISDA 2018)*, Vellore, India, Springer International Publishing, 2020, vol. 940, pp. 138–146.

[10] K. Cong *et al.*, "KGDetector: Detecting chinese sensitive information via knowledge graph-enhanced BERT," *Secur. Commun. Netw.*, vol. 2022, pp. 1–9, 2022. doi: 10.1155/2022/4656837.

[11] A. Mehdi *et al.*, "A brief survey of text mining: Classification, clustering and extraction techniques," 2017, *arXiv:1707.02919*.

[12] J. Y. Hou, "Research on deep learning-based detection technology for sensitive text information in scene images," (in Chinese), North University, China, 2021. doi: 10.27470/d.cnki.ghbgc.2021.001033.

[13] W. Y. Li, "Research and application of detection technology for sensitive information in images based on global feature elements," (in Chinese), University of Chinese Academy of Sciences (Institute of Computing Technology, Chinese Academy of Sciences, Shenyang), China, 2022. doi: 10.27587/d.cnki.gksjs.2022.000019.

[14] T. Ashwini and C. Caragea, "Image privacy prediction using deep neural networks," *ACM Trans. Web*, vol. 14, no. 2, pp. 1–32, 2020. doi: 10.1145/3386082.

[15] J. L. Shi and X. F. Zhao, "Anti-leakage method of network sensitive information data based on homomorphic encryption," *J. Intell. Syst.*, vol. 32, no. 1, 2023, Art. no. 20220281. doi: 10.1515/jisys-2022-0281.

[16] A. Kaul, M. Kesarwani, H. Min, and Q. Zhang, "Knowledge & learning-based adaptable system for sensitive information identification and handling," in *2021 IEEE 14th Int. Conf. Cloud Comput. (CLOUD)*, Chicago, IL, USA, 2021, pp. 261–271.

[17] L. Gao, X. Wu, J. Wu, X. Xie, L. Qiu and L. Sun, "Sensitive image information recognition model of network community based on content text," in *2021 IEEE Sixth Int. Conf. Data Sci. Cyberspace (DSC)*, Shenzhen, China, 2021, pp. 47–52.

[18] Y. Mao, B. Song, Z. Zhang, W. Yang, and Y. Lan, "A lightweight image sensitive information detection model based on yolov5s," *Academic J. Comput. Inform. Sci.*, vol. 6, no. 3, pp. 20–27, 2023.

[19] D. Hu, X. Hou, L. Wei, L. Jiang, and Y. Mo, "MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, Singapore, May 23–27, 2022, pp. 7037–7041.

[20] X. Yan, H. Xue, S. Jiang, and Z. Liu, "Multimodal sentiment analysis using multi-tensor fusion network with cross-modal modeling," *Appl. Artif. Intell.*, vol. 36, no. 1, 2022, Art. no. 2000688. doi: 10.1080/08839514.2021.2000688.

[21] S. Zou, X. Huang, X. Shen, and H. Liu, "Improving multimodal fusion with Main Modal Transformer for emotion recognition in conversation," *Knowl.-Based Syst.*, vol. 258, 2022, Art. no. 109978. doi: 10.1016/j.knosys.2022.109978.

[22] P. Y. Ji *et al.*, "Adaptive sensitive information recognition based on multimodal information inference in social networks," *Secur. Commun. Netw.*, vol. 2023, no. 1, 2023, Art. no. 5627246. doi: 10.1155/2023/5627246.

[23] X. Gao, J. Yu, Y. Chang, H. Wang, and J. Fan, "Checking only when it is necessary: Enabling integrity auditing based on the keyword with sensitive information privacy for encrypted cloud data," *IEEE Trans. Dependable Secur. Comput.*, vol. 19, no. 6, pp. 3774–3789, Nov. 1–Dec. 2022. doi: 10.1109/TDSC.2021.3106780.

[24] J. H. Wang, Y. T. Wu, and L. Wang, "Predicting implicit user preferences with multimodal feature fusion for similar user recommendation in social media," *Appl. Sci.*, vol. 11, no. 3, 2021, Art. no. 1064. doi: 10.3390/app11031064.

[25] S. G. Xiao and W. P. Fu, "Visual relationship detection with multimodal fusion and reasoning," *Sensors*, vol. 22, no. 20, 2022, Art. no. 7918. doi: 10.3390/s22207918.

[26] M. Zhang and R. Chunara, "Leveraging CLIP for inferring sensitive information and improving model fairness," 2024, *arXiv:2403.10624*.

[27] B. Xu *et al.*, "Different data, different modalities! reinforced data splitting for effective multimodal information extraction from social media posts," in *Int. Conf. Comput. Linguist., 2022*, Gyeongju, Republic of Korea, Oct. 12–17, 2022.