2022 19th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP) | 978-1-6654-9389-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCWAMTIP56608.2022.10016592

DEEP LEARNING BASED SENSITIVE DATA DETECTION

PENG CHONG[1]

[1]School Of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610000, China
E-MAIL: 202021081128@std.uestc.edu.cn

Abstract:

The growing popularity of edge techniques, such as IoT, 5G, blockchain, make it increasingly challenging to protect sensitive data due to the amount of data increases and the growing volume of regulatory policies. To properly protect sensitive data, it is very important to identify sensitive data and implement data anonymization to ensure the quality and proper use of data anonymization techniques. This work focuses on proactively sensitive data identification, classification and anonymization using machine learning techniques. We first investigated the sensitive data extraction from both structured data and unstructured data, in which Bert models and Regular expressions were used to achieve the identification of sensitive data in real-time. Meanwhile, we propose a comprehensive sensitive detection framework combining the Bert model with regular expressions that can achieve high precision and good generalization capability with not so large corpus. The experimental results demonstrate the effectiveness of proposed solution.

Keywords:

Sensitive data detection; Data anonymization; Deep Learning; Cyber intelligence

1. Introduction

The advances in edging computing technologies, e.g., 5G, Internet of Things (IoT), cloud computing, big data, artificial intelligence (AI), etc. are transforming our lives in remarkable ways and propelling social and economic development [1]. However, while these new technologies bring convenience to our life, they also inevitably make data more vulnerable to infringement and causes more serious consequences. It is also quite easy to immediately access users' personal individual information (PII) through third-party applications or software, which may result in privacy exposure [3]. Leakage of such sensitive information will disrupt users' daily lives, create incalculable losses, and even endanger their personal safety. It is clear that privacy detection and protection has always been one of the most important research directions [4]. For example, Jalaluddin proposed an lightweight cosine function for encryption methodology to protect personal patients' informative data [2].

The machine learning shows great potential in sensitive information protection. Machine learning is a technology that allows computer systems to gain expertise and anticipate or recognize unknown facts by training them on large amounts of data, which is making significant influence in a variety of fields, such as medical care, biology, industry IoT, agriculture, food, robots, finance, and other fields all benefit from it's high accuracy rate and strong generalization capabilities [5]. One of the most common uses of information extraction in machine learning is NER (Named Entity Recognition), which involves extracting entities from text such as person, location, and organization. NER cannot only operate as an independent and mature system, but also as the foundation for a variety of complicated natural language processing tasks such as response systems and text summarization [6]. Although machine learning works well, large sensitive data sets and complex machine learning models need to be built when dealing with a wide variety of sensitive data. Sometimes it is easier and more convenient to use a simple rule-based approach.

In this paper, we propose a comprehensive sensitive information detection method for generalized personal privacy information, which is classified into two categories: privacy information with structured features and privacy information with unstructured features, using regular expressions and machine learning methods, respectively, to achieve high accuracy and high recall rate.

The main contributions of this work are summarized.

1) In this work, a sensitive data detection and identification framework are proposed based on machine learning and regular expressions, which is able to extract both structured and unstructured sensitive data.
2) A novel sensitive data detection method is proposed which combines regular expressions with a machine learning model to attain excellent accuracy and recall rate.

978-1-6654-9389-5/22/$31.00 ©2022 IEEE

Authorized licensed use limited to: Hochschule Rosenheim (FH). Downloaded on November 01,2023 at 13:27:01 UTC from IEEE Xplore. Restrictions apply.

2. Related Works

2.1. Sensitive Information and Privacy Detection

Sensitive personal information usually refers to personal information such as bio-metrics, religious beliefs, specific identity, medical and health information, financial accounts, track, and other information that, if leaked or illegally used, can easily result in a violation of personal dignity or harm to personal and property safety [7].

Using machine learning model (e.g., deep learning model) to detect and protect privacy data is a novel and challenging research area. The machine learning approach focuses on building comprehensive and complex NER systems to detect sensitive information. Paulo Silva trained and evaluated data sets containing personally identifiable information using three well-known natural language processing tools (NLTK, Stanford, and CoreNLP) [8]. Adeyemo Victor Elijah developed an intrusion detection system with LSTM model which can achieve a detection accuracy rate of 80% on the two-classed attack dataset [9]. Compared with the traditional detection model, deep learning can automatically discover potential rules. It can also achieve high accuracy and is able to guarantee generalization ability. The deep learning-based privacy detection shows great promises in privacy detection protection.

2.2. Deep learning based named entity recognition

Extract specific information is one of the most important application of deep learning. In recent years, the deep learning-based NER model has become the mainstream and has produced the most advanced results. In contrast to feature-based approaches, deep learning can automatically discover potential representations and features that is required for classification or detection [10].

The key to deep learning is to train word vectors using various neural network structure models. Collobert et al. employed a CNN (convolution neural network) to produce local features around each word and input them into the label decoder to compute the distribution score of potential labels after each word in the input sequence was embedded into the N × N dimension vectors [11]. Using the gate mechanism, the LSTM model can avoid it and perform well in extended sequences. Chalapathy et al. achieve 85.19 F1-score (under an unofficial evaluation) on MedLine test data adding a CRF layer to the top of a LSTM model [12].

Transformer has been a huge success in many areas of artificial intelligence, such as natural language processing, computer vision and audio processing. In comparison to circulating neural networks such as RNN and LSTM, the Transformer structure's self-attention mechanism can parallelize the amount of computation, overcoming the limitation that RNN and LSTM models cannot do. Transformer can also build more interpretable models cause each attentional head may learn to execute different tasks. BERT which stacked by lots of transformer encoder blocks achieved the best results of 11 NLP tasks at the time [13].

3. Deep Learning Based Comprehensive Sensitive Data Detection Framework

In this section, this paper focus on how we build a comprehensive sensitive data detection system. In Section 3.1, this paper focus on the definition and classification of sensitive information and private data. In Section 3.2, this paper gives an introduction of the overall comprehensive sensitive data detection framework. In Section 3.3, this paper focus on how we detected structured sensitive data using regular expression. Finally in Section 3.4, this paper focus on how we detect unstructured sensitive data using machine learning.

3.1. Sensitive Information and Private Data

Sensitive information is the information that individuals, institutions do not want to be known to the outside world. In specific applications, sensitive information is mainly related to personally identifiable information, patient illness records, company financial information, etc. Desensitization mainly refers to the reliable protection of sensitive data through the desensitization of these sensitive information. In order to accurately evaluate the effect of desensitization, this study further subdivides private data into structured privacy data and unstructured privacy data.

In data anonymization, the definition of personal data is unclear yet but general following personal data is considered 'sensitive' and is subject to scenarios: personal data, health records, ethic, religion/political/sex opinion, etc.

3.1.1 Definition of Sensitive Information

. *Definition 1. Static sensitive information*: Sensitive information ($R_{u_i} = r^1_{U_i}, r^2_{U_i}, ..., r^R_{U_i}$) that has a fixed structure for a particular user u_i. Each rule $r^k_{U_i}$ is a privacy rule with a specific structure associated with the user.

$$\otimes \leftarrow f_1 \wedge f_2 \wedge ... \wedge f_L \quad (1)$$

in which $\otimes$ means the target sensitive rule r_i. the right side of the (1) is the description of regular logical operation. Each f_k is represented as a logical expression for an instance property. L mainly indicate the length of the rule. The

Authorized licensed use limited to: Hochschule Rosenheim (FH). Downloaded on November 01,2023 at 13:27:01 UTC from IEEE Xplore. Restrictions apply.

expression of the equation is

$$IF f_1 \& f_2 \& ... \& f_L, THEN Class = \otimes \quad (2)$$

In order to detect R_u from user data, there are several methods. For simpler rules, one can use regular expressions to extract the rules directly. For the data with complex features, the method based on machine learning can be used to automatically detect and identify the features.

3.1.2 Classification of Sensitive Information

As mentioned above, in general, Rules based privacy detection can be divided into two groups: text privacy data and visual privacy data. For text privacy data, it can be further categorized into structured data and unstructured data.

Definition 2. Structured sensitive data: the rules of the sensitive data are regular and structured that can be expressed easily by regular expressions.

Definition 3. Unstructured sensitive data: The rules of this sensitive data are irregular. Its rules are hard to express directly. For the detection of such data, this paper adopts the method of machine learning to mine the hidden rules or patterns.

According to the above definition. This research lists 11 kinds of sensitive information with private rules in life. The sensitive information is shown in Table 1.

Table 1 Typical Sensitive Data Categories

Structured Data	Email address; Phone Number; Passport; Ip Address
Unstructured Data	Person; Location; Organization; Disease; Occupation

3.2. Comprehensive Sensitive Data Detection Framework

Automatically sensitive data detection and classification play a key role in data anonymization. In this section, a sensitive data automatically detection framework will be developed, which is able to detect sensitive features from sensitive data which is consisted of structured data and unstructured data. The framework also classifies these sensitive data according to their subjects. The sensitive data automatically detection framework's key procedure can be shown in Fig.1.

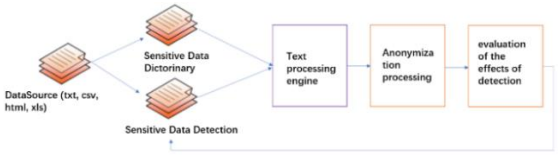

Fig.1 Procedure of Detection Framework

Data Source As defined in Section 2, unstructured data typically features with flexible formats, these unstructured data are embedded in different sources, such as *texts, documents, weblogs, images, etc.* In this work, we focus on unstructured sensitive data in text, such as *name, address* and so on.

Sensitive Data Detection In this research, a comprehensive Detection module is proposed to deal with sensitive data. In this module, we will detect the structured sensitive data firstly with the regular expression in Section 3.2. Then for the unstructured sensitive data, we adopt machine Learning technology. Machine Learning technology that can automatically classify sensitive data/files can significantly reduce the risk of exposure of sensitive data or based on the detected result, we can alert publisher on potential sensitive data.

Text processing engine Once sensitive data has been detected and identified, it will be automatically tagged sensitive tags, such as *PII, commercial sensitive, etc.* These tags usually can be created by specific user based on their scenarios. We will also create Sensitive Data Dictionary to store these sensitive tags and data. Using Sensitive Data Dictionary, we can search and encrypt sensitive data more conveniently.

Anonymization and processing the identified sensitive data will be anonymized using specific algorithms. In this research, we just encrypt these sensitive data with their class. if necessary, we can adopt better algorithms including k-anonymous, l-diversity, etc. to protect identified sensitive data.

Evaluation of the effects of detection We analyzed the results of the detection from three perspectives. For structured data, we analyze the advantages and disadvantages of regular expression. For unstructured data, we analyze our detection model's params and F_1 score. The experiment proves that the accuracy is at the cost of the computation

3.3. Sensitive Information Detection from Structured Data

It is very important to detect and identify sensitive data before protecting them. The key to the detection of structured sensitive data is to identify the structure patterns. In this

Authorized licensed use limited to: Hochschule Rosenheim (FH). Downloaded on November 01,2023 at 13:27:01 UTC from IEEE Xplore. Restrictions apply.

subsection, this paper will introduce features of structured sensitive data in daily life and study the corresponding regular expressions to detect them.

Email address is an important personal data. Leaked email address may be the target of scammers and spammers or even worse endanger personal property and life safety. To protect your email address, it needs to be identified and anonymized in some public scenarios. Using regular expression ^[a-zA-Z0-9_.-]+@[a-zA-Z0-9-]+a-zA-Z0-9-.] can detect email address.

Telephone number is the most commonly used data in life. Once your phone number is leaked. you may face countless harassment messages which will cause great trouble to your life. Using regular expression 1(?:[358][0-9]|4[579]|66|+7[0135678]|9[89])[0-9]{8}+ to detect phone number.

Passport is one of the necessary documents for going abroad. It is a legal document to prove the nationality and identity of the citizen. In China, passports are divided into diplomatic passports, official passports, ordinary passports and special zone passports. Using regular expression ^(E\d{8}) | E[A-Za-z]\d{7} | (G\d{8}) | (H\d{8}) | (HJ\d{7}) | (K\d{8}) | (KJ\d{7}) | (MB\d{7}) can detect passport.

IP address is assigned by Internet service provider, which is the network address of device in the Internet. IP address is very important for keeping us safe. Hackers are able to conduct cyber attacks if they found specific target IP address. In this study, we only detect common IPv4 addresses. IPv4 addresses are typically made up of four groups of numbers, each ranging from 0 to 255. Using regular expression ((?:[0,1]?\d{1,2}|2(?:[0-4][0-9]|5[0-5]))(?+:\.(?:[0,1]?\d{1,2}|2(?:[0-4][0-9]|5[0-6]))) +{3})+ to detect ip address.

It is not very challenging to detect and identify structured sensitive data since fixed patterns can be used. Actually, most sensitive data are embedded in unstructured data, e.g., documents, images, audio, etc., which needs sophisticated techniques, such as machine learning to analyze.

3.4. Unstructured Sensitive Data Detection using Machine Learning

Unstructured sensitive data usually features flexible formats which makes it hard to detect with regular expressions. This section focuses on how to detect unstructured sensitive data with machine learning methods. In practice, a appropriate model should be selected into the comprehensive sensitive data detection framework.

3.4.1 Dataset

This paper creates a specific privacy dataset for training and testing because there are few personal privacy datasets on the Internet. The private dataset we needed was scattered across other datasets. Therefore, this paper extracts the sensitive data from other datasets to construct the final privacy dataset of sensitive information. From the famous Conll-2003 datasets [14], we extracted 6000 pieces of data including PERSON, LOCATION and ORGANIZATION. From the public NCBI-Disease dataset [16], this paper extracted 2000 pieces of data containing disease information. Datasets about Occupation entities are scarce on the Internet. So, this paper extracted 2000 pieces of occupation sentences from CLUE Chinese dataset [17], translated them into English and merged them into our final sensitive dataset.

Table **2.** shows the contents in the final sensitive dataset.

Table 2 Contents of Sensitive DataSets

DataSet	Description
Conll-2003 DataSet	Sentences that contains name, location, and organization
CLUE DataSet	Sentences that contains occupation tags
NCBI DataSet	Sentences that contain common disease tags

3.4.2 BERT

The sequential serial computing process of LSTMs greatly increase the cost of computing because the LSTMs calculation must be performed after the completion of the previous moment. However, BERT network uses Attention mechanism [15] instead of RNNs to make the computation parallel, thus greatly reducing the computation cost. In addition, BERT is an excellent transfer learning model. Through pre-training in a large number of unsupervised expectations, BERT learns the deep-language feature representation of contextual information. Pre-train and fine-tune make BERT achieve the best results in 11 NLP tasks [13].

In our research, we use BERT network to detect private data. The architecture of BERT uses a series of Transformer blocks which contain self-attention mechanism, stacked on top of each other. Each transformer block takes word embeddings as input which are constructed by the encoding of the word vector. Considering the data scale of our current study, we adopted the standard of Bert-Base parameter. In our BERT model, we adopt 12 stacked Transformer blocks, each with a feed-forward network containing 768 hidden units and 12 attention heads. And the input of our model

Authorized licensed use limited to: Hochschule Rosenheim (FH). Downloaded on November 01,2023 at 13:27:01 UTC from IEEE Xplore. Restrictions apply.

takes in less 512 word-vectors at a time.

After the transformer block, BERT is fine-tuned for a specific task. In our research, this task is named entity recognition. So, we add a linear layer with SoftMax classifier to score every token to indicate the most probable entity. After the linear layer, we also add a CRF layer similarly to learn the transition rule to make the result better. The whole structure of our model can be shown in Fig.2.

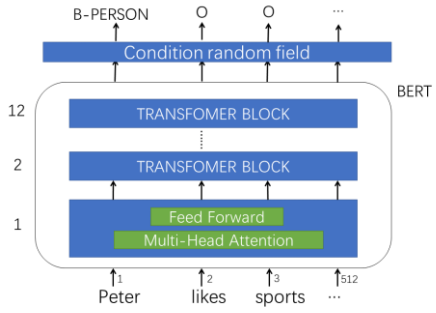

Fig.2 The architectural of BERT model

4. Experimental Validation

4.1. Analysis of experimental results

our sensitive data detection system can automatically detect the sensitive information and encrypt them with the corresponding type. For the Structured Data Detection, our system can recognize 4 classes as described in Table 1.

Besides the structured Data, our system can also detect and encrypt the unstructured data, the unstructured data usually features with flexible formats, such as name, locations, address, etc. For the Unstructured Data Detection, our system can recognize 5 classes as described in Table 1.

Fig.3. shows our system's detection results for the Sensitive Data.

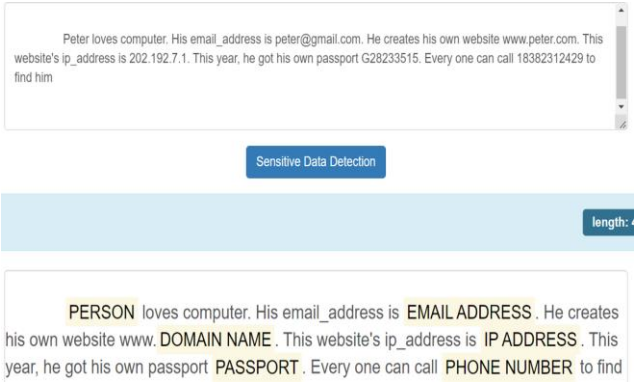

Fig.3 an example of sensitive data detection.

4.2 Experiments Results Analysis

4.2.1 Machine Learning results Analysis

F1 score is a comprehensive measure of accuracy and recall rate. It accurately represents the performance of model training. It's the harmonic mean of the models' Precision and Recall computations. The formula is as follows:

$$F_1 = 2 * \frac{precision*recall}{precision+recall} \quad (3)$$

Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while recall (also known as sensitivity) is the fraction of relevant instances that were retrieved. there are 1272 sensitive data in our test dataset. The specific value can be shown in Table 3.

Model params is usually used to compare the computational complexity of model which allows comparison of Computation without regard to Hardware. The higher the value of param, the complexes the model is. In our Bert model, every input char will be represented into a 512-shape vector and there are 768 hidden units for computing the weights.

Table **3**. shows the exact param number and a comprehensive comparison between these three models.

Table 3 Performances Summary

Model	Precision	Recall	F_1 score	Params
BERT	0.897	0.896	0.896	110M
Regular Expressions	1.000	\	\	\
BERT+Regular expressions	0.925	0.896	0.896	110M

In Table 3, we can learn that combing BERT with regular expressions to detect sensitive result we can get good result.

4.2.2 Comprehensive Detection Framework Analysis

Traditional sensitive data detection mainly relies on specific rules such as regular expressions described in Section 3.2. The advantage of traditional detection method is high precision, but it is difficult to build complete rules or dictionaries since some sensitive data is flexible and irregular. Also, the traditional detection method has low generalization capability which makes it hard to detect to new sensitive data. Machine Learning method not only has well generalization capability, but also has high precision. However, Machine Learning method needs extremely large corpus to build it's datasets. Thus, Obviously, The Comprehensive Detection Framework that defining the structured sensitive data and the

Authorized licensed use limited to: Hochschule Rosenheim (FH). Downloaded on November 01,2023 at 13:27:01 UTC from IEEE Xplore. Restrictions apply.

unstructured data, combining regular expressions with BERT model can achieve high precision and good generalization capability with not so large corpus.

5. Conclusions

This work focused on the sensitive data proactively identification and anonymization using machine learning based techniques. Specifically, this work investigated the sensitive data extraction techniques from structured data and unstructured data, in which a machine learning based sensitive detection framework was proposed that can automate the identification of sensitive data in real-time with deep learning model BERT. The proposed method can achieve 92.5% precision and 89.6% recall rate without so large corpus.

Acknowledgements

I would like to thank my supervisor (shancang.li) for his tireless guidance and dedication. Without him, this article would not have happened.

References

[1] H.-Y.Tran and J.Hu, "Privacy-preserving big data analytics a comprehensive survey," Journal of Parallel and Distributed Computing, vol.134, pp.207-218, 2019

[2] J.Khan, G.A.Kan, J.P.Li, et al."Secure smart healthcare monitoring in industrial internet of things(iiot) ecosystem with cosine function hybrid chaotic map encryption," Scientific Programming, vol. 2022,2022.

[3] A. De Slave, P.Mori, and L. Ricci, "A survey on privacy in decentralized online social networks," Computer Science Review, vol.27,pp.154-176,2018.

[4] D.K.Alferidah and N.Jhanjhi, "Cybersecurity impact over big data and iot growth," in 2020 International Conference on Computational Intelligence(ICCI). IEEE,2020,pp.103-108.

[5] V.Meshram, K.Patil, V.Meshram, D.Haanchate, and S.Ramkteke, "Machine learning in agriculture domain:a state-of-art survey," Artificial Intelligence in the Life Sciences, vol.1,p.100010,2021.

[6] J.Li,A.Sun,J.Han, and C.Li,"A survey on deep learning for named entity recognition," IEEE Transactions on Knowledge and Data Engineering,vol.34,no.1,pp.50-70,202.

[7] Y.-I.Liu, L.Huang, W.Yan, X.Wang, and R.Zhang, "Privacy in ai and the iot: The privacy concerns of smart speaker users and the personal information protection law in China," Telecommunications Policy, vol.46, no.7, p.102334, 2022

[8] P.Silva, C.Goncalves, C.Godinho, N.Antunes, and M.Curado, "Using nlp and machine learning to detect data privacy violations," in IEEE INFOCOM 2020-IEEE conference on Computer Communications Workshops(INFOCOM WKSHPS). IEEE, 2020,pp.972-977.

[9] V.E.Adeyemo, A.Abdullah, N.JhanJhi, M.Supramaniam, and A.O.Balogun, "Ensemble and deep-learning methods for two-class and multi-attack anomaly intrusion detection: an empirical study," International Journal of Advanced Computer Science and Applications,vol.10,no.9,2019.

[10] J.Ma,J.Zhang,L.Xiao,K.Chen,"Classification of power quality disturbances via deep learning," IEEE Technical Review, vol.34, no.4,pp.408-415,2017

[11] R.Collobert and J.Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in Proceedings of the 25th international conference on Machine learning, 2008, pp.160-167.

[12] R.Chalapathy, E.Z.Borzeshi, and M.Piccardi, "An investigation of recurrent neural architectures for drug name recognition," arXiv preprint arXiv:1609.07585, 2016.

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

[14] E.F.Sang and F.De Meulder, "Introduction to the conll-2003 shared task: Language-independent named entity recognition,"arXiv preprint cs/0306050,2003.

[15] A.Vaswani, N.Shazeer, N.Parmar, J.Uszkoreit, L.Jones, A.N.Gomez,L.Kaiser, and I.Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol.30,2017.

[16] R.I.Dogan, R.Leaman, and Z.lu, "Ncbi disease corpus: a resource for disease name recognition and concept normalization," Journal of biomedical informatics, vol.47,pp.1-10,2014

[17] Xu L,Zhang X,Li L, et al. CLUE:A Chinese Language Understanding Evaluation Benchmark.2020.

Authorized licensed use limited to: Hochschule Rosenheim (FH). Downloaded on November 01,2023 at 13:27:01 UTC from IEEE Xplore. Restrictions apply.