Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

Internet and Telecommunication Fraud Prevention Analysis based on Deep Learning

Peifeng Ni & Quanxiu Wang

To cite this article: Peifeng Ni & Quanxiu Wang (2022) Internet and Telecommunication Fraud Prevention Analysis based on Deep Learning, Applied Artificial Intelligence, 36:1, 2137630, DOI: 10.1080/08839514.2022.2137630

To link to this article: https://doi.org/10.1080/08839514.2022.2137630

Published online: 03 Nov 2022.

APPLIED ARTIFICIAL INTELLIGENCE
2022, VOL. 36, NO. 1, e2137630 (3425 pages)
https://doi.org/10.1080/08839514.2022.2137630

Taylor & Francis
Taylor & Francis Group

Internet and Telecommunication Fraud Prevention Analysis based on Deep Learning

Peifeng Ni[a] and Quanxiu Wang[b]

[a]School of Economics and Management, University of Chinese Academy of Sciences, Beijing, P.R, China; [b]AI Research, RICH AI, Beijing, China

ABSTRACT

In recent years, contactless fraud crimes via telecommunication and Internet have grown rapidly. Meanwhile, the rate of solved criminal cases is much lower, which is mainly due to two reasons. Firstly, the definition of risk factors in the field of new Internet and telecommunication fraud crime is not comprehensive, resulting in the problem not being well defined. Secondly, Internet fraud crime information is mostly recorded using natural language with huge volume, and there is a lack of automated and intelligent way to deeply analyze and extract the risk factor. To better analyze the Internet and telecommunication fraud crime to help solve more cases, in this paper, we propose a new Internet and telecommunication fraud crime risk factor extraction system. After studying the existing related research, we propose a novel risk factor extraction technology based on BERT. This novel technology can gracefully deal with multi-sources and heterogeneous data problems during the extraction of risk factors in multiple dimensions; meanwhile, it can significantly reduce the need for computation resources and improve the online serving performance. After experimentation, this technique can significantly reduce training time by 60%-70%, and meanwhile, it can reduce the computation resources by 80% and improve serving performance by 5 times during serving. In our approach, we propose a novel approach to set sample weight and loss weight based on data characteristics and data distribution during model training, which can significantly improve extraction precision. With adjusting the sample weight during model training, we can get 1.56% precision improved. Moreover, setting the loss weight during model training, the precision can be improved by 1.63% compared to baseline mode.

ARTICLE HISTORY
Received 24 August 2022
Accepted 13 October 2022

Introduction

Risk is a systematic and multidimensional concept which corresponds to the word "security." The popularity of the Internet has improved work productivity, but also brought new space and opportunities to conduct illegal and criminal activities. Compared to traditional fraud, the newly emerging Internet and telecommunication fraud is to utilize the Internet or

CONTACT Peifeng Ni, nipeifeng19@mails.ucas.ac.cn School of Economics and Management, University of Chinese Academy of Sciences, No.19(A) Yuquan Road, Shijingshan District, Beijing, P.R 100049, China

telecommunication as medium to conduct fraud, including telephone, SMS, WeChat, QQ, and other telecommunications network platforms. This new type of non-contact crime can easily and quickly reach large populations, which brings new risks to the society and causes huge loss to the victims.

In the past 10 years, the number of Internet telecommunication fraud crimes has increased at a rate of 20%–30%. Reports from the related government department show that the number of Internet and telecommunication fraud crime cases filed in 2015 was 590,000, causing losses of 22 billion RMB; the number of Internet and telecommunication fraud crime cases filed in 2016 was 630,000, causing losses of 19.7 billion RMB; the number of Internet and telecommunication fraud crime cases filed in 2017 was 864,000, causing property losses of 17.07 billion RMB. Due to the huge impact on society and the public, the public security authorities have launched a series of dedicated campaigns to resolve the problems. In 2017, the national public security authorities solved 78,000 cases of Internet and telecommunication fraud crimes and arrested 47,000 criminals; in 2018, the national public security authorities have solved 131,000 cases of Internet and telecommunication fraud crimes and arrested 73,000 criminals, recovered 2.03 billion RMB of losses, and prevented 9.7 billion RMB from fraud. Although the number of cases solved has increased year over year, the solving rate is still relatively low compared to the huge number of cases filed (Tencent Guardian 2017a, 2017b, 2018).

Despite the rapid growth of Internet and telecommunication fraud crimes, there are not a lot of relevant analyses to help deeply understand Internet and telecommunication fraud crimes and to prevent such crimes. This is mainly due to two reasons. First, the definition of Internet and telecommunication fraud crime risk factors has not been well defined. Second, the Internet and telecommunication fraud crime information is mostly recorded using natural language in a textual form, with huge volume and diverse forms, while the existing risk factor extraction capability can mostly only deal with single-factor type extraction or at most several types of factor extraction. This brings challenges to the analysis, prediction, and early warning of new types of Internet and telecommunication frauds.

In this paper, we propose a novel Internet and telecommunication fraud crime risk factor extraction technology based on BERT (Devlin et al. 2019) to help better analyze the Internet and telecommunication fraud crime. This novel technology can gracefully deal with multi-sources and heterogeneous data problems and can extract multiple risk factors at the same time. This technology can also significantly reduce the need for computation resources and improve the online serving performance. In this approach, we propose a novel technique to set sample weight and loss weight based on data characteristics and data distribution during model training, which can significantly improve extraction precision and recall. Based on the proposed technology, we

can implement an automatic extraction system to extract risk factors from huge volume of Internet and telecommunication fraud cases, which can further help the analysis, prediction, and early warning of new types of Internet frauds.

The contribution of this paper is as follows:

(1) We proposed a comprehensive Internet and telecommunication fraud crime risk factor knowledge framework by applying Fault Tree Analysis method based on the existing research and domain experts' knowledge.
(2) We constructed a BERT-based fusion technique to solve the problem of heterogeneous risk factor extraction from multiple data sources. Compared to traditional methods, this novel approach can significantly reduce the training time by 60%–70%. During serving, this technique can also reduce the computation resources by 80% and improve the serving performance by 5 times.
(3) Considering that different data sources may have different amounts of labeled data and different probabilities of occurrence, we proposed a novel technique to adjust the sample weight during model training, which improves the precision by 1.56% comparing to the traditional random sampling method.
(4) Considering the difference of different risk factors, we proposed a method to adjust loss weight during model training to improve the weight of uncommon risk factors in the loss calculation, which improves precision by 1.63% comparing to models trained without loss weight adjustment.

The paper is structured as follows: in Section 2 we present related work, in Section 3 describes the Internet and telecommunication fraud risk factor knowledge framework, in Section 4 describes our methodology includes the model and experimental methods, in Section 5 describes the datasets used, the experiments and the results analysis, and finally, we conclude our work and outline future work in Section 6.

Related research

After conducting the literature survey, we found only a few studies have been conducted to analyze the risk factors of Internet and telecommunication fraud crimes. Due to that most Internet fraud crimes are described and recorded in textual natural language, if we want to conduct comprehensive and accurate analysis of risk factors without human involved, the system must understand what is contained in the textual information, and need to parse and comprehend the textual descriptions, and then convert the information contained in

the text into structured representation so that it can be further analyzed. Traditional risk factor extraction methods include the following types:

(1) Risk factor extraction method based on traditional word statistics. This method uses vector to represent the textual content, where each element in the vector indicates the frequency of each word appearing in the textual content. It then divides the textual content into two parts: the part containing common information and the part containing valuable and differential information. These words are then ordered according to the word statistics method, such as TF-IDF (Salton, Yang, and Yu 1975), word frequency (Luhn 1957), word co-occurrence (Matsuo and Ishizuka 2003), word lexicality, and so on. These statistics represent the importance of words in the content, and finally calculate the risk factors based on these statistics. The advantages of traditional word statistics-based methods are simple and easy to apply with low computational requirements, but they have many drawbacks such as poor applicability, loss of low-frequency words, and inability to explore new risk factors.
(2) Risk factor extraction method based on topic model. This method first uses topic models to divide the textual content into different topics, and then uses different methods to extract key information on each topic as the risk factors. For example, Yijun Gu et al. proposed a method of extracting risk factors by applying TextRank (Mihalcea and Tarau 2004) based on the application of LDA (Blei, Ng, and Jordan 2003) topic model. However, such methods are prone to extract many words that are weakly associated with risk factors, leading to bias in risk factor extraction.
(3) Risk factor extraction methods based on named entity recognition as a sequence labelingtask(Huang, Xu, and Yu 2015; Lample et al., 2016; Ma and Hovy, 2016). Recent progress in named entity recognition has evolved from lexicon and rule-based methods to traditional machine learning methods, and to deep learning-based methods. For example, The BiLSTM-CRF (Huang, Xu, and Yu 2015) model, which uses a combination of neural networks and traditional methods, can improve semantic parsing capability, with the limitation that it still cannot fully utilize the context due to the network structure. Since Google introduced the BERT (Devlin et al. 2019) network, a new trend to use BERT-CRF to solve the named entity recognition problem has quickly become popular (He, Chen, and Wen 2022; Liu et al. 2021). The BERT network uses Attention (Vaswani et al. 2017) technology, transformer network, and MaskLM tagging method, which can realize the bidirectional semantic encoding of the full text. For Multi-task training, Qian Chen et al. proposed a joint intent classification andslot

filling model based on BERT(Chen, Zhuo and Wang 2019) which increase improvement onintent classification accuracy and slot filling F1 in the same dataset.

Based on the recent advances in technology, it is not difficult to extract one or several types of entities, and it's also not difficult to combine different training tasks in the same dataset, but in the scenario of Internet and telecommunication fraud risk factor analysis, the existing entity extraction models cannot solve the problem of multi-sources and heterogeneous data very well, mainly due to the following reasons.

(1) In case only one model can be applied to extract risk factor, traditional model training requires consistent training data labeling methods for all datasets. While for the new types of Internet and telecommunication fraud crimes, the risk factors involve a wide range of dimensions, usually including hundreds of factors, and different datasets may focus on different risk factors, which brings challenges in data annotation and labeling, and consequently leads to low training data quality and poor model performance.
(2) In case multiple models can be applied to extract risk factor, as the amount of risk factor labels increases, the number of models also increases, and the training cost increases too. Meanwhile, a large amount of computation resources will be required during online model inference. To complete a single task, we must assemble multiple inferences, which also leads to serving performance degradation.
(3) Because the new types of Internet and telecommunication fraud crimes are evolving extremely fast, the risk factor knowledge framework also needs to be updated frequently. Such definition updating requires the extraction model to be updated, which may require re-labeling of all training datasets to cover the additional or updated risk factors, which leads to a significant increase in training efforts.

Therefore, there is an urgent need to design better models which can handle multi-sources and heterogeneous data to reduce the efforts of data labeling, reduce the training cost, and improve the inference performance.

Internet and Telecommunication Fraud Risk Factor Knowledge Framework

The characteristics of Internet and telecommunication fraud and the composition of risk factors can be studied in several dimensions, including the tools to conduct crimes, modus operandi, targets of crime, distribution of victim geography, and crime organization.

(1) Analysis of the fraudulent tools: Internet and telecommunication fraud takes various forms and evolves extremely quickly. By the end of December 2020, the number of China's Internet users reached 989 million and the Internet penetration rate is 70.4%; the number of cell phone users reached 986 million, and the proportion of China's Internet users using cell phones to access the Internet is as high as 99.70% (CNNIC(China Internet Network Information Center) 2021). The commonly used instant messaging applications are extremely easy to be used by criminals to conduct fraud. According to the Cybercrime Judicial Big Data and Ten Typical Cases of Telecom and Network Fraud Crimes released by The Supreme People's Court of The People's Republic of China (2019), online messaging platforms such as WeChat, QQ, TikTok, and so on, have become the most frequently used medium to conduct Internet and telecommunication fraud crimes, which has led to a chain of gray industry. In recent years, criminals have also infiltrated their crimes into more online platforms, such as Animal World, Transaction Cat, QingXinYiGou, Firecoin, and so on. Considering the fast growth of new types of Internet and telecommunication frauds, it is important to automatically discover the new Internet fraud platforms to help identify fraud risks.
(2) Analysis of modus operandi: The Internet and telecommunication fraud pattern is constantly evolving, the fraud scripts are updated very quickly, and the fraudulent techniques of Internet fraud crimes are constantly updated according to the awareness of the public and the prevention. Early fraudulent patterns such as "pretending to be a government agent," "guess who I am," "winning a grand prize on a TV program," and so on, are gradually becoming well known, and the success rate of such types of fraud becomes lower and lower, so such fraudulent patterns were used less and less common. Meanwhile, the Internet and telecommunication fraud gangs have developed new fraudulent patterns such as naked chat, part-time job, express delivery, investment, and so on, which are closely tying to the daily lives of targeted victims. According to the published articles by public security departments, the type of Internet and telecommunication fraud can be roughly divided into 48 categories, including more than 300 subcategories (Cybercrime Judicial Big Data and Ten Typical Cases of Telecom and Network Fraud Crimes 2019). The new types of fraudulent means and patterns of criminals emerge fast, endlessly, and are also more covert and disorienting. Therefore, how to automatically discover the new criminal patterns is another key to fraud risk identification, and it needs to be incorporated into the new Internet and telecommunication fraud risk factor analysis framework.

(3) Targeted audience analysis: To improve the success rate, criminals cast a wide net, involving a wide range of geographic areas and many audiences. Depending on the fraudulent techniques, the selected groups are also different. For example, in student loans fraud, fraudsters often select college students as the target; in health products fraud, fraudsters often select the elderly as the target. Therefore, the analysis of targeted audience is important to help precisely identify and prevent Internet and telecommunication fraud risks.
(4) Analysis of victim geography: according to the "Anti-telecommunication Internet and telecommunication fraud big data report in the first quarter of 2018," Guangdong province ranks first in the number of fraudulent phone calls, text messages, and virus infections, in addition to Shandong, Jiangsu, Sichuan, Henan, and Guangxi (Li 2017). After studying the published data from 2017 to 2019, the Internet and telecommunication fraud crimes have been concentrated in the provinces of Zhejiang, Guangdong, Fujian, Henan, Jiangsu, Anhui, and Hainan. Therefore, the analysis of the geographic area of Internet fraud victimization is beneficial to the precise prevention of Internet fraud crime risks (Shi 2021).
(5) Analysis of fraudulent property: With the diverse forms of Internet and telecommunication fraud, fraudulent property has developed from the initial single type (money) to multiple forms, such as virtual property, including game accounts, game equipment, virtual currency, and so on. In addition to bank transfers, the transfer methods have also developed into a variety of channels, including red envelopes, sweep codes, WeChat transfers, platform transfers, and other forms. The diversity of fraudulent property and transfer methods also makes it more difficult to block and stop payments.

In this paper, we proposed a comprehensive Internet and telecommunication fraud crime risk factor knowledge framework by applying Fault Tree Analysis method (Vesely et al. 1981) based on the existing research and domain experts' knowledge. In this approach, the new Internet and telecommunication fraud risk is regarded as a "fault" generated in the process of social development, the Internet and telecommunication fraud risk generation process of "fraudulent tools," "modus operandi," "targeted audience," "victim geography," "fraudulent property," "payment method," are the "parts" which lead to system failure. According to the logical relationship of the fault tree from the system to the components, and then to the parts, according to the "descending" analysis method, we can build a fault tree diagram which displayed in Figure 1.

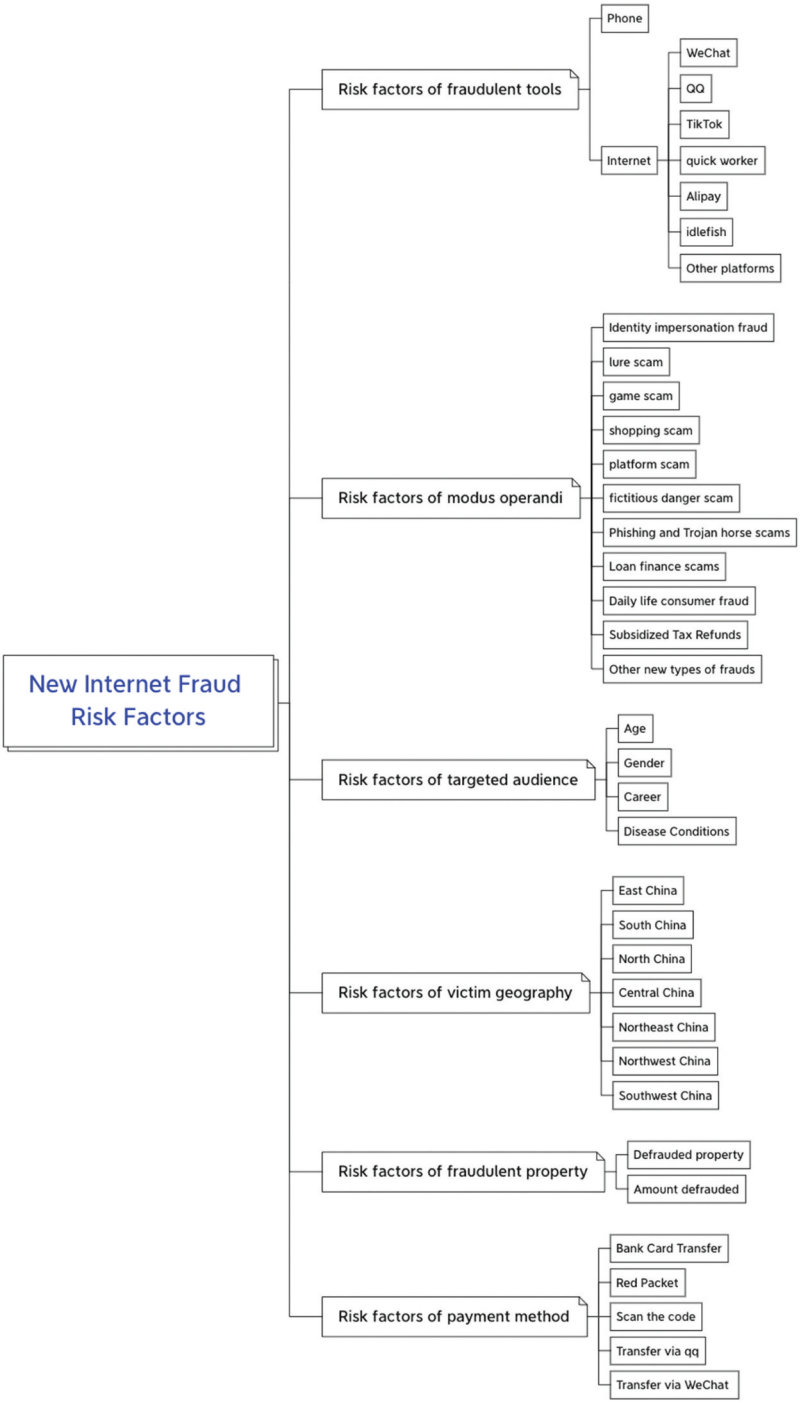

Figure 1. New internet and telecommunication fraud risk factors.

Model and Experimental Methods

BERT-Based Fusion Factor Extraction Model for Multi-Sources and Heterogeneous Datasets

To overcome the problems of the complex and quickly evolving risk factor knowledge framework, the current technology needs large amount of labeled training data and huge model training workload. This results in the inability to dynamically adapt the system and update the data, and so on. To solve the Internet and telecommunication fraud risk factor analysis, in this paper, we propose to use BERT-based fusion extraction technology to solve the problem of factor extraction from multi-sources and heterogeneous data. The input data supports multiple data sources, and different data sources can be labeled heterogeneously with different factor labels.

The main part of the model adopts the structure of 1+N, where 1 is the BERT body shared by multiple datasets, and N is the number of risk factor labels corresponding to N classifiers. This model structure supports the horizontal expansion of datasets and avoids repeated feature calculation and data labeling. The model structure is displayed in Figure 2.

The training process is as follows:

Step 1: for each batch, sample *batch_size* of data from multiple datasets.

Step 2: encode data, where in this paper, we use BIOES tagging scheme to encode each category of factors. We also convert the categories of risk factors labeled in each data into a mask, which will be used when calculating the loss function.

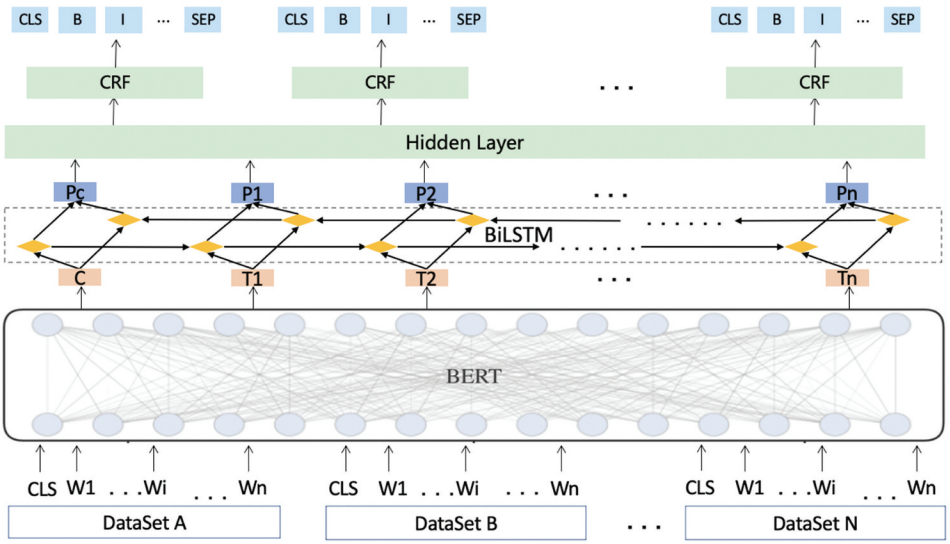

Figure 2. Model structure.

Step 3: use BERT as the main framework of the fusion network structure, and pass the character information and location information of the text data into BERT, where BERT will encode the representation of them.

Step 4: input vector sequences of sentences generated by BERT into two stacked bidirectional LSTM layers, apply batch normalization between these two layers, and process the text for sequence representation, and then feed the obtained representation sequences, again, into a hidden layer composed of a fully connected network.

Step 5: In the output layer, different risk factors correspond to different CRFs, and separate gradient updates are performed for different factor labels. Further transfer matrix parameters are introduced to make the final sequence with bounded relationships, so that the start span and end span of the BIOES labels can be predicted more accurately, and the risk factor types can be decoded.

Step 6: In the decoding stage, after calculating the loss of all risk factor types of each data, multiply it with the "mask of which factor types are labeled in each data" in step 2, so that the unlabeled factor types do not participate in the optimization of the loss.

This BERT-based fusion extraction model can conduct different risk factor extraction tasks based on the same BERT pre-training model, which considers both the commonality of the model and the characteristics of each task; meanwhile, the risk factor training data can be accumulated from multiple datasets covering different factor types and different text types. Advantages of this novel approach include:

(1) Multiple heterogeneous datasets with different labels can be combined in one model to help the model improve accuracy, which greatly improves data utilization.
(2) It can avoid repeated labeling of data and reduce labeling cost.
(3) When model inference is made, only one model needs to be deployed, i.e., multiple results can be predicted at the same time, reducing deployment and hardware costs.
(4) It is possible to train both public datasets and private datasets, thus improving the model's capability of generalization.

Sample Weight

Considering that different risk factors may have different amount of labeled data and different probabilities of occurrence in different datasets, we modify the data selection process in the training stage by adding *sample_weight*, which controls the sampling ratio of different datasets in each batch. This method improves the possibility of seeing more diversity labeled information in the same batch training. In this paper, we refine the random sample strategy by

smoothing, which raises the sampling ratio of factor types with less data in the training process, thus improving the accuracy of factor type in less data.

The data sampling process based on the addition of *sample_weight* is as follows.

Step 1: Randomize the order of data in each dataset.

Step 2: Give each dataset a pointer, initially pointing to the first piece of data in each dataset.

Step 3: Select a dataset according to *sample_weight*. For example, there are currently three datasets, dataset1, dataset2, and dataset3, each with a sample weight of 0.3, 0.5, 0.2, then sampling with a random probability between (0,1). Select dataset2 when the probability is between [0.0, 0.3), select dataset2 when the probability is between [0.3, 0.8), and select dataset3 when the probability is between [0.8, 1.0).

Step 4: Fetch the data currently pointed to by the pointer in that dataset, and the pointer is moved back one position. When the pointer has moved to the end, then the dataset is randomly disordered again, and the pointer points to the first data.

Step 5: Repeat steps 3 and 4 each time when getting the data.

We set *sample_weight* configuration, respectively, for the random sample and smooth sample strategy as follows:

Random Sample

$$sample\_weight_i = \frac{d_i}{\sum_{j=1}^{n} d_j} \tag{1}$$

where ***n*** denotes how many datasets there are, and d_i denotes the amount of data in the *i-th* dataset. The random *sample_weight* is related to the amount of dataset.

Smooth Sample

$$smooth\_sample\_weight_i = \min\left(\log\frac{\sum_{j=1}^{n} d_j}{d_i}, 20\right) * d_i \tag{2}$$

$$sample\_weight_i = \frac{smooth\_sample\_weight_i}{\sum_{j=1}^{n} smooth\_sample\_weight_j} \tag{3}$$

where ***n*** denotes how many datasets there are, and d_i denotes the amount of data in the *i-th* dataset. $smooth\_sample\_weight_i$ is the sampling weights of the *i-th* dataset with smoothing before normalization. $sample\_weight_i$ is the final sampling weights of the *i-th* dataset after normalization. In this way, the factor types in less data will be sampled more frequently.

Loss Weight

The main body of the model adopts a 1+N structure, with N corresponding to N classifiers, each of which is responsible for the serial labeling of one risk factor. Considering the differentiation of risk factors, such as the factor types with lower probability of occurrence, the model is more difficult to learn, so compared to the model of calculating the loss of each data equally, this paper compared the loss function without adjustment and the loss function with smooth loss weight, and proposes the loss weight of fusion factor which integrates two influencing factors: one is the labeling quantity of each factor type, and the other is annotations appearing frequency of the factor type in the dataset which it belongs. This method will increase the weight of uncommon risk factors in the loss calculation.

Three types of loss functions are set as follows:

Baseline Loss Formula Without Adjustment

$$loss = -\sum_{l=1}^{M} y_l log(p_l) \tag{4}$$

where $\boldsymbol{M}$ is the amount of factor types.

Loss Formula with Smooth Loss Weight

$$\widehat{loss\_weight_l} = min\left(\log \frac{\sum_{j=1}^{M} e_j}{e_l}, 20\right) \tag{5}$$

$$loss\_weight_l = \frac{\widehat{loss\_weight_i}}{\sum_{j=1}^{M} \widehat{loss\_weight_j}} \tag{6}$$

$$loss = -\sum_{l=1}^{M} y_l log(p_l) * loss\_weight_l \tag{7}$$

where $\boldsymbol{M}$ is the amount of factor types, e_l denotes the total amount of data for the *l-th* factor type, $loss\_weight_l$ denotes the *l-th* factor type's loss weight, and we give less weight to more data in the training process. ***loss*** is the final loss obtained by applying the loss weight to the original loss.

Loss Formula with Loss Weight of Fusion Factor

$$factor\_show\_out\_rate_l = \frac{e_l}{d_l} \tag{8}$$

$$\widehat{loss\_weight_l} = min\left(\log\frac{\sum_{j=1}^{M} e_j}{e_l}, 20\right) * \frac{1}{factor\_show\_out\_rate_l} \tag{9}$$

$$loss\_weight_l = \frac{\widehat{loss\_weight_i}}{\sum_{j=1}^{M} \widehat{loss\_weight_j}} \tag{10}$$

$$loss = -\sum_{l=1}^{M} y_l log(p_l) * loss\_weight_l \tag{11}$$

where $\boldsymbol{M}$ is the amount of factor types, e_l denotes the total amount of data for the *l-th* factor type, d_l denotes the sum of the data volume of those datasets containing the *l-th* factor type. We use $factor\_show\_out\_rate_l$ to denote the occurrence probability of *l-th* factor. The observation is that the lower the value, the more difficult it is to learn in the training process. We apply the $factor\_show\_out\_rate_l$ to generate $loss\_weight_l$. ***loss*** is the final loss obtained by applying the loss weight to the original loss.

Experiments and Results Analysis

Data Collection

In this paper, the experiment data are from police notification data obtained from the Internet, as well as fraud case from media data and case document data from multiple judgment websites. After the screening of fraud-related keywords, a total of 3504 texts are obtained. The data distribution and labeling are displayed in Table 1.

All the obtained datasets were annotated, and the overall distribution of risk factors in the four datasets are displayed in Table 2.

Table 1. Data distribution and labeling.

Dataset	Description	Data labeling status	Number of data items
dataset1	Police notification data obtained from Internet channels-Part1	The dataset is labeled with 12 types of tags such as fraudulent tools and modus operandi	436
dataset2	Police notification data obtained from Internet channels-Part2	The dataset is labeled with 11 types of tags such as fraudulent tools, victim geography, fraudulent property	1294
dataset3	Fraud case from media data	The dataset is labeled with 19 types of tags such as modus operandi, targeted audience, and payment method	1166
dataset4	Case document data from multiple judgement website	The dataset is labeled with 23 types of tags such as fraudulent tools, modus operandi, targeted audience, victim geography, and fraudulent property	608
Total			3504

Table 2. The labeled data distribution of risk factor.

Risk Factor Category	Risk factor level 2 category	Data volume
Risk factors of fraudulent tools	Phone	367
	Network	1241
Risk factors of modus operandi	Identity impersonation scam techniques	81
	Lure-type fraudulent practices	249
	Game type fraudulent techniques	169
	Shopping scams	85
	Fictitious risk scamming techniques	22
	Phishing, Trojan horse-type fraudulent practices	83
	Loan financial fraudulent practices	349
	Daily life consumption fraudulent techniques	180
	Subsidized tax refund scam techniques	27
	Other new types of fraudulent practices	323
Risk factors of targeted audience	Age	74
	Gender	188
	Career	48
	Disease Conditions	12
Risk factors of victim geography	East China	140
	South China	127
	North China	90
	Central China	132
	Northeast China	81
	Northwest China	72
	Southwest China	104
Risk factors of fraudulent property	Defrauded property	202
	Amount defrauded	665
Risk factors of payment method	Bank Card Transfer	169
	Red Envelop	47
	Scan the code	29
	QQ transfer	43
	WeChat transfer	87

Evaluation Criteria

To verify the resource usage and serving performance of the model when in use, we use the following metrics for comparison.

(1) Hardware resource *trainingResource_GPU* used by the model during training, GPU usage time *trainingTime_GPU* during training, and training span time *training_duration*.
(2) The average GPU resources occupied by the model inference service, including idle GPU resource *inferenceResource_GPU_idle*, busy GPU resource *inferenceResource_GPU_busy*.
(3) *DPS*: the number of texts that can be processed per second (full amount of fraud risk factors extracted from the texts).

Meanwhile, to verify the performance of the model on the test set, the accuracy (P), recall (R), and F1 used were selected in this study and calculated as shown in Equation (12) to Equation (14).

$$P = \frac{TP}{TP + FP} \tag{12}$$

$$R = \frac{TP}{TP + FN} \tag{13}$$

$$F1 = \frac{2PP}{P + R} \tag{14}$$

In the formula, TP (True Positive) denotes the number of risk factors correctly identified; FP (False Positive) denotes the number of incorrectly identified risk factors; FN (False Negative) denotes the number of manually labeled risk factors that are not correctly identified by the model.

Performance Analysis

The experiments are based on CPU i9-9900K and GPU GeForce RTX 2080Ti as the training test environment. The parameters of the network structure are displayed in Table 3.

The following two comparison experiments are conducted with fixed parameters.

(1) Baseline experiment: without the fusion extraction method of this paper, the optimal method to deal with the datasets in this paper, is to train five different risk factor extraction models to complete different training tasks. The 5 models training tasks are displayed in Table 4.

(2) Fusion Model Experiment: Only 1 Fusion Model Needs to Be Trained to Complete All Risk Factor Extraction Tasks. The model training task is displayed in Table 5.

Table 3. Parameter settings.

Parameter	Value	Parameter	Value
maximum sequence length	256	learning_rate	0.00005
batch_size	16	epoch	20
drop_out	0.1	warmup_proportion	0.1
num_tuning_layers	12	num_BiLSTM_layers	2
Optimizer	AdamW	Activation function	GELU

Table 4. Baseline experiment.

Model	Task	Data source
Model 1	Training in extraction of factors of fraudulent tools	dataset1, dataset2, dataset4
Model 2	Training in extraction of factors of modus operandi	dataset1, dataset3, dataset4
Model 3	Training in the extraction of factors of targeted audience	dataset3, dataset4
Model 4	Training in extraction of factors of victim geography, fraudulent property	dataset2, dataset4
Model 5	Training in extraction of factors of payment method	dataset3

Table 5. Fusion model experiment.

Model	Task	Data source
Fusion Model	Training in extraction of fraudulent tools, modus operandi, targeted audience, victim geography, fraudulent property, and payment method	dataset1, dataset2, dataset3, dataset4

The relevant resource occupancy and performance during the training process are displayed in Table 6.

The relevant resource occupancy and service performance during model inference are displayed in Table 7.

Therefore, in terms of model design, BERT-based fusion factor extraction technique is constructed to solve the problem of handling multi-sources and heterogeneous datasets, which significantly reduces hardware resources and improves usage performance compared to non-fused models. On the dataset involved in this paper, the training time is reduced by 60%–70% in the training phase with the same hardware resources, and in the model inference phase, the hardware resources are saved by 70–80% and the performance is improved by nearly 5 times.

Metrics Analysis

Sample_weight Effect Comparison

To test the data sampling weight proposed in this paper, this paper compares the experimental effects of two *sample_weight*, and trains two BERT-based fusion factor extraction models, respectively, on the same training dataset, and calculates the P, R, and F1 metrics of the two models on the same test data, respectively. The metrics are compared in Table 8:

The tensorboard graph is shown in Figure 3

Table 6. Resource usage during training.

trainingResource_GPU	Experiment	*trainingTime_GPU*	*training_duration*
One GPU, 11 GB of video memory	baseline experiments (Serial execution of 5 training tasks)	92 min	92 min
	Fusion experiments (Only 1 training mission required)	29 min	29 min
	Time optimization ratio	68.5%	68.5%
5 GPUs, 55 GB of video memory	baseline experiments (Parallel 5 training tasks）	94 min	26 min
	Fusion experiments (Can be accelerated with multi-card parallel training)	37 min	8 min
	Time optimization ratio	60.6%	69.2%

Table 7. Resource usage during model inference.

Experiment	*inferenceResource_GPU_idle*	*inferenceResource_GPU_busy*	*DPS*
baseline experiments (5 models)	7875 MB	8991 MB	48
Fusion experiments (1 model)	1575 MB	2691 MB	235
Resource optimization ratio	80%	70.10%	-

Table 8. Effect comparison.

Experiment	precision	recall	f1	Increase
random sample (formula 1)	80.44%	76.95%	78.66%	-
smooth sample (formula 2–3)	81.25%	79.20%	80.21%	1.56%

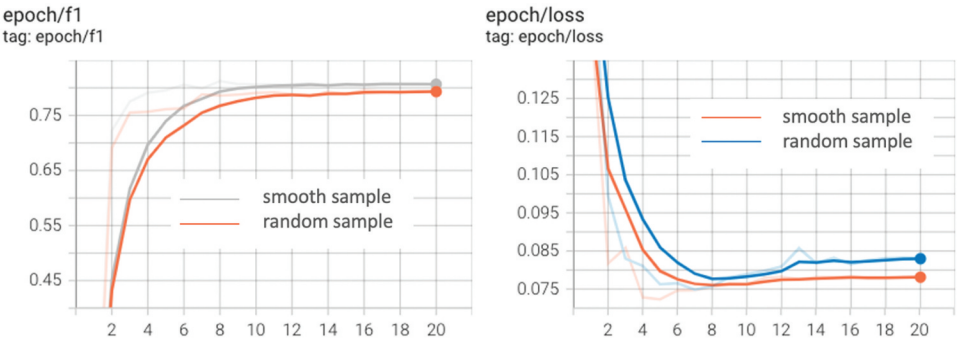

Figure 3. Tensorboard graph.

Therefore, in considering risk factors with different amount of labeled data and different probabilities of occurrence, the smooth sample weight is increased to improve the accuracy of factor types with less data, and the overall accuracy is improved by 1.56% compared to the random sampling method.

Loss_weight Effect Comparison

To test the effectiveness of the loss setting method for the probability of occurrence of fusion factors proposed in this paper, this paper compares the experimental effects of three loss formulas by training three BERT-based fusion factor extraction models on the same training dataset, and calculating the P, R, and F1 metrics of the three models on the same test data, respectively. The metrics are compared in Table 9:

The tensorboard graph is shown in Figure 4:

Table 9. Effect comparison.

	precision	recall	f1	Increase
baseline loss formula without adjustment (formula 4)	81.62%	77.68%	79.60%	-
loss formula with smooth loss weight (formula 5–7)	82.71%	77.71%	80.13%	0.53%
loss formula with loss weight of fusion factor (formula 8–11)	83.73%	78.88%	81.23%	1.63%

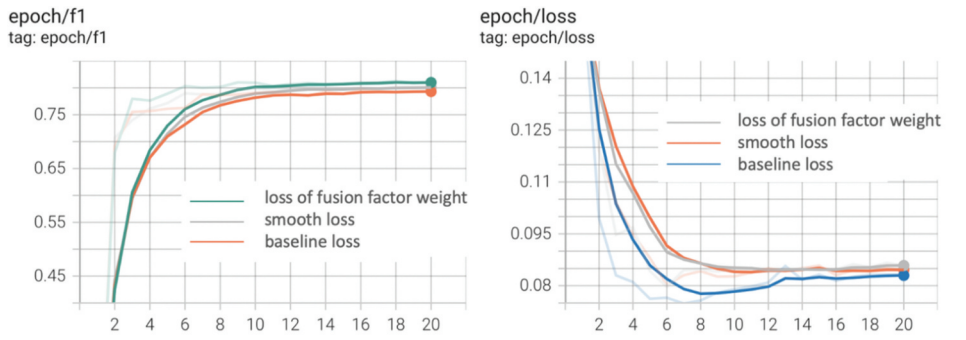

Figure 4. Tensorboard graph.

Therefore, in considering the differences of different risk factors, we increase the loss weight of the probability weight of occurrence of fused risk factors in the loss calculation and increase the weight of uncommon risk factors in the loss calculation, which improves the accuracy by 1.63% compared with the experiment of baseline.

Conclusion and Future Work

The new Internet and telecommunication fraud risk factors proposed in this paper define the new Internet fraud risk system in multiple dimensions, and the feasibility and effectiveness of the methods in this paper are verified by combining the police notification data from Internet, fraud case from media data and case document data from multiple judgment websites. The BERT-based fusion factor extraction technique for multi-sources and heterogeneous datasets adopted in this paper significantly reduces hardware resources, shortens the training time, and improves the model inference performance compared with the non-fusion model; the smooth data sampling method and the loss method with fused factor weights adopted in this paper effectively improve the extraction accuracy of the new Internet and telecommunication fraud-based risk factor system.

In this paper, the extraction of risk factors is elaborated, but the early warning prediction based on risk factors is not fully analyzed. How to model the early warning prediction based on risk factors is an area worthy of further study, which will be part of the future work.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Chen, Q., Z. Zhuo, and W.Wang. 2019. Bert for joint intent classification and slot filling. arXiv preprint arXiv:1902.10909.

CNNIC (China Internet Network Information Center). 2021. The 47th China Statistical Report on Internet Development, China Internet Network Information Center, China. Accessed February 03. http://cnnic.cn/hlwfzyj/hlwxzbg/hlwtjbg/202102/P020210203334633480104.pdf

Devlin, J., M. W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota.

He, T., J. Chen, and Y. Wen. 2022. Research on entity recognition of electronic medical record based on BERT-CRF model. *Computer & Digital Engineering* 50(3):639–643 .

Huang, Z., W. Xu, and K. Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv*:1508.01991.

Lample, G., M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. 2016. Neural architectures for named entity recognition. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 260–270, San Diego, California. Association for Computational Linguistics.

Li, J. 2017. On legal countermeasures to curb cybercrime. *Network Security Technology & Application*. Issue 1:2–4.

Liu, X., M. Zhang, Q. Gu, Y. Ren, D. He, and W. Gao. 2021. Named entity recognition of fresh egg supply chain based on BERT-CRF architecture. *Transactions of the Chinese Society for Agricultural Machinery*. doi:10.6041/j.issn.1000-1298.2021.S0.066.

Luhn, H. P. 1957.A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development* 1: 309–17.doi: 10.1147/rd.14.0309

Ma, X.Z. and E. Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1064–1074, Berlin, Germany Volume 1: Long Papers. Association for Computational Linguistics.

Matsuo, Y., and M. Ishizuka. 2003. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools* 13:157–69. doi:10.1142/S0218213004001466.

Mihalcea, R., and P. Tarau. 2004. TextRank: bringing order into texts. *Proceedings of the 2004 conference on empirical methods in natural language processing*:404–411.

Salton, G., C. S. Yang, and C. T. Yu. 1975.A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science* 26: 33–44.doi: 10.1002/asi.4630260106

Shi, Z. X. Research on the governance path of internet fraud crime from the perspective of risk society. *Journal of Chinese People's Armed Police Force Academy*. Issue 4: 33–39 2021. April

The Supreme People's Court of The People's Republic of China. 2019. Cybercrime judicial big data and ten typical cases of telecom and network fraud crimes. Accessed November 19, 2019. https://www.court.gov.cn/zixun-xiangqing-201181.html

Tencent Guardian, 2017a. Big data report on anti-telecom and internet fraud in the second quarter of 2017. Accessed August 04, 2017a. https://tg110.qq.com/single.html

Tencent Guardian, 2017b. Big data report on anti-telecom and internet fraud in the third quarter of 2017. Accessed November 13, 2017b. https://tg110.qq.com/newspage/report_center_18_1_26.html

Tencent Guardian, 2018. Big data report on anti-telecom network fraud in the fourth quarter of 2017. Accessed February 08, 2018. https://tg110.qq.com/newspage/report_center_20180208page1.html

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems. doi:10.48550/arXiv.1706.03762.

Vesely, W. E., F. F. Goldberg, N. H. Roberts, and D. F. Haasl. 1981. *Fault tree handbook (NUREG-0492)*. Washington, DC: U.S. Nuclear Regulatory Commission.