

Roberta-MHARC: Enhanced Telecom Fraud Detection with Multi-head Attention and Residual Connection

Jun Li¹^[0000-0001-5591-721X] and Cheng Zhang²^[0009-0009-4694-5900]

School of Computer Science and Information Security,
Guilin University Of Electronic Technology, Guilin 541000, China
lijun5221@126.com

Abstract. Telecom fraud has become one of the hardest-hit areas in the criminal field. With the development of artificial intelligence technology, telecom fraud texts have become highly concealed and deceptive. Existing prevention methods, such as mobile phone number tracking, detection, and traditional machine learning model text recognition, lack real-time performance in identifying telecom fraud. Furthermore, due to the scarcity of Chinese telecom fraud text data, the accuracy of recognizing Chinese telecom fraud text is not high. In this paper, we design a telecom fraud text decision-making model Roberta-MHARC based on Roberta combined with multi-head attention mechanism and residual connection. First, the model selects some categories of data in the CCL2023 telecom network fraud data set as basic samples, and combines it with the collected telecom fraud text data to form a five-category covering impersonating customer service, impersonating leadership acquaintances, loans, public security fraud, and normal text data set. Secondly, during the training process, the model adds a multi-head attention mechanism and improves the training speed through residual connections. Finally, the model improves its accuracy on multi-classification tasks by introducing an inconsistency loss function alongside the cross-entropy loss. Experimental results show that our model achieves good results on multiple benchmark datasets.

Keywords: natural language processing, telecom fraud, Roberta, multi-head attention, multi-classification tasks.

1 Introduction

With the rapid advancement of the information society, cybercrimes, particularly telecom fraud, have emerged as predominant criminal activities globally, growing at an alarming rate. According to a collaborative study conducted by the Global Anti-Fraud Alliance and data service provider ScamAdviser, between August 2022 and August 2023, approximately 25.5% of the world's population fell victim to telecom network fraud, resulting in losses surpassing \$1 trillion. Telecom network fraud has thus become a formidable challenge for global law enforcement efforts [1]. In China, the National Anti-Fraud Center has taken decisive actions to combat this menace. It issued a staggering 9.406 million fund warning instructions, engaged with 13.89 million individuals

to dissuade them from falling prey to fraudulent schemes, collaborated with relevant agencies to intercept a staggering 2.75 billion fraudulent calls and 2.28 billion text messages, and addressed 8.364 million domain names and URLs associated with fraud. Furthermore, emergency interception of funds involved a substantial amount totaling 328.8 billion yuan [2]. These proactive measures underscore China's commitment to tackling telecom fraud and safeguarding its citizens from financial exploitation in the digital realm.

In order to combat telecom fraud, academia, governments, and technology companies around the world are making efforts to this end. The academic community has proposed [3-7] numerous models to detect telecom fraud, and [8] also proposed a public dataset Fake-base-station helps academics conduct research. In China, the government has launched the National Anti-Fraud Center app to help people prevent telecom fraud. Some companies are also doing some work to prevent telecom fraud. For example, Tencent has developed the Tencent Technology Anti-Fraud Mini Program, which aims to help raise the national anti-fraud awareness, report telecom fraud crimes, and help people in a more convenient way. Users can reduce losses and provide them with a series of guidance on telecom network fraud. Despite significant progress, telecom fraud remains severe both domestically and internationally. For instance, criminals may exploit new events occurring every year. For example, the new coronavirus incident in 2019 has led to many telecom fraud incidents. However, the models and datasets designed previously lack such data, making mistakes easy to occur. Therefore, a more comprehensive data set and predictive telecom fraud model that is more in line with the current social form are needed.

In natural language processing fraud detection models, such as [9-12], results have been achieved in some fields, such as credit card financial fraud, medical fraud, e-commerce fraud, etc. These models have achieved good results in their respective fields, and then when applying these powerful models to telecom fraud, because the behavior of telecom fraud is different from the above-mentioned fraud methods, the performance results are poor. A key problem is that telecom fraud mainly relies on text-based behaviors, which is different from the behavior of text-based fraud. The other three types of fraud are different. Therefore, when telecom fraud is so rampant, a model that can detect it well is urgently needed. This study aims to address these pressing needs for the application of natural language processing in telecom fraud. In this work, we propose a natural language processing model for telecom fraud aimed at identifying and classifying telecom fraud texts. Our biggest work lies in the Roberta-MHARC (Residual Roberta with Multi-Head Attention) model, which is a model specially designed to handle telecom fraud texts. The following are the salient features and contributions of this study:

1. Designs a telecom fraud model capable of identifying and classifying fraudulent information accurately in text. The Chinese telecommunications fraud model we proposed is a relatively rare research direction at present.

2. Based on the latest data, constructs a Chinese 5-category dataset covering impersonating customer service, impersonating leadership acquaintances, loans, public security fraud, and normal text. It fills the gap in telecom fraud in Chinese datasets and

provides help to other researchers, allowing our research to better reflect and solve current social problems.

3. Our exploration and application of the inconsistency loss function and the cross-entropy function. Although these two loss functions were used separately in previous research, we innovatively combined them to form a new loss function framework and conducted a large number of experiments to verify its effectiveness in the field of telecommunications fraud. The uniqueness and effectiveness of this integrated approach enable our model to achieve significant progress in fraud detection, providing new ideas and methods for solving this important social problem.

Together, these innovations make the Roberta-MHARC model as a powerful weapon in the fight against telecom fraud. By improving detection accuracy, we hope to greatly reduce the occurrence of fraud and create a safe telecom environment. The remainder of this paper is organized as follows: Section 2 provides a review of existing literature on the application of NLP techniques for telecom fraud assistance. Section 3 describes the modules used in this NLP model in detail. Section 4 shows the data set used in this study and the collection and construction process, as well as the experimental results, including performance indicators. Finally, Section 5 summarizes the full paper and proposes future research directions.

2 Related Work

Telecom fraud detection is a constantly updated and iterative task, intertwined with various special events. In this increasingly complex situation, sophisticated analysis and understanding tools are required.

2.1 Introduction to neural network model

Natural language processing technology has become a leading representative in this battle, promising to extract key information from raw, unstructured text. In early research, the main method relied on compiling blacklists of fraudulent phone numbers. For example, Emmanuel [13] and others proposed a content-based method to detect telecom fraud. This method can effectively limit the situation of continuous fraud with the same number. However, this approach often comes into play after the fraud has already taken shape. Criminals typically exploit the same number to send out widespread information and then engage in targeted fraudulent activities. By the time victims report it, the fraud has usually already occurred, without a timely and effective preventive effect. There are also rule-based studies, such as Qianqian Zhao [14] using natural language processing to extract features from text data. They further detect telecom fraud by establishing rules to identify similar content within the same call. There are also methods based on machine learning. For example, Ileberi [15] proposed a method that combines random forest (RF), decision tree (DT), artificial neural network (ANN), naive Bayes (NB) and logistic regression (LR) feature selection method based on genetic algorithm (GA). Dornadula [16] proposed a fraud detection model based on genetic algorithm, which also used random forest and adaboost algorithm. At the same

time, in order to alleviate the problem of class size imbalance, the smote sampling method was also used. These methods played a certain role in the early days when the amount of data was small and the telecom infrastructure services were average. However, with the advent of the information age and the big data era, these methods are no longer sufficient to solve the task of telecom fraud detection. Based on pre-training Models and neural network models can show more powerful advantages. These models have the advantage of learning representations directly from text without the need for feature engineering. The cutting-edge of these models include the recurrent neural network (RNN) [17], the long short-term memory network (LSTM) [18], the gated recurrent unit (GRU) [19] and attention mechanisms [20], which have received much attention due to their outstanding performance in NLP tasks. The structure of the RNN model itself is very suitable for processing data with time series characteristics. The state at any moment is related to the state at the previous moment, and the text data context is just related. When the number of network layers increases, the RNN model is prone to the problem of gradient disappearance or gradient explosion. This leads to the long short-term memory network that is improved from RNN. The input gate, forgetting gate and output gate are unique in LSTM. Three gate structures, these gate structures have functions similar to brain cells, and can selectively store some information or forget some information. Each gate structure controls the amount of information flowing in the model, thereby alleviating the gradient disappearance and gradient explosion problems caused by long-term recording sequence relationships. Jianbing Yan [21] designed an intelligent algorithm for text classification based on convolutional neural network (CNN) and bidirectional long short-term memory network (BiLSTM). There is another improved variant, gated recurrent unit. Unlike LSTM, the gate structure is simplified, and update gates and reset gates are used to control the flow of information. Li X [22] used GRU to form a model stacked by an ensemble model, a deep sequence learning model and a top-level ensemble classifier, and achieved very good results in the field of transaction fraud. The attention mechanism allows the model to automatically focus on the most important parts when processing sequence data, thereby improving the performance of the model. The attention mechanism can enhance the feature weight of fraud texts and fraud-related keywords, and can also adapt to fraud texts of different lengths. Javier [4] proposed a method for detecting fraud based on the Transformer model. They modeled the user's spending profile and detected fraud behaviors that deviated from it. They also used attention-based mechanisms and unsupervised learning to further improve the accuracy of the model. Meng Z [23] proposed a credit card fraud detection method based on generative adversarial network (GAN) and multi-head attention mechanism. Zhou J [24] proposed an online fraud text recognition method based on recurrent neural network, multi-head attention mechanism and convolution. The multi-head attention mechanism is used to increase the model's ability to learn global interactive information and multi-granularity local interactive information in online fraud texts. In short, the attention mechanism can better understand and identify telecommunications fraud texts.

2.2 Pre-trained model

With the proposal of the attention mechanism, large-scale pre-training models also emerged. For example, BERT [25] introduced a bidirectional transformer encoder to pre-train a large-scale language model, enabling the model to deeply understand the semantics in the context and grammatical relations, it uses the Masked Language Model (MLM) pre-training task, in which the model needs to predict the masked content of some tokens in the input sequence, thereby forcing the model to understand the contextual relations in the sentence. The success of BERT has also inspired many variants, such as Roberta [26], DistilBERT [27], etc., which have been improved and customized in various tasks and application fields. Xinze Yang [28] proposed the FinChain-BERT model, which improves the ability to process complex financial text information through deep learning technology. The negative log likelihood function and the keyword function were compared in the loss function. The result showed that the keyword The loss function is better. Jishnu K S [29] introduced a phishing URL detection method, using Roberta to extract semantic and contextual information from URLs. In this literature description, the work of telecommunications fraud detection is far from over. People need more powerful, efficient, and accurate models to help people better face the telecom fraud crisis. Therefore, we propose the Roberta-MHARC model.

3 Methodology

3.1 Roberta

Roberta (Robustly optimized BERT approach) is a pre-trained language model based on the Transformer architecture. Improved on BERT, Roberta has achieved significant success in the field of natural language processing. During the calculation we need to convert the text data into word vectors. Traditional word vector representations include bag-of-words models, TF-IDF representations, Word2Vec and GloVe. The relatively new word vector model now includes BERT to build word vectors. BERT is a pre-trained language model based on the transformer architecture and achieved great success in the field of natural language processing. However, Roberta performs well in multiple natural language processing tasks, usually better than BERT. This may be attributed to larger training data and better training strategies. Roberta improved performance by using larger models, longer training time, removing NSP (Next Sentence Prediction) tasks, using dynamic masks and other improvements. In this study, we use Roberta to perform word embedding processing on text to obtain rich semantic information.

3.2 Multi-head attention mechanism

Multi-head attention [20] is a key technology used in deep learning to improve the model's ability to process sequence data. By using multiple attention heads simultaneously, multi-head attention allows the model to learn and capture different relationships

and patterns in different attention subspaces. This improves the model's ability to express input sequences, especially in fields such as natural language processing. The proposal of multi-head attention greatly enriches the expressive ability of the attention mechanism, allowing the model to pay more flexible attention to multiple aspects of the input sequence, helping to deal with complex sequence relationships. In this study, a 12-head attention mechanism was used. The following is the calculation formula of the multi-head attention mechanism:

$$\text{MultiHead}(Q,K,V)=\text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_{12})W^O \quad (1)$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), i = 1, \dots, 12 \quad (2)$$

Among them, Q , K , and V represent the query, key, and value respectively, W_i^Q , W_i^K , and W_i^V represent the corresponding weight matrix respectively, W^O represents the output weight matrix, and head_i represents the output of the i -th head.

3.3 Residual connection

Another way to solve the problem of gradient disappearance and gradient explosion is residual connection [30]. The core idea of residual connection is to pass the input directly to the output by adding a direct path that bypasses part of the network. The benefit of this design is that even if some layers learn identity mappings, the overall network can still learn more complex nonlinear mappings. Residual connections help gradients flow more smoothly through the network during backpropagation, mitigating the vanishing and exploding gradient problems, making it easier to train and optimize deeper networks. The formula for residual connection is as follows:

$$x_{l+1} = x_l + F(x_l, W_l) \quad (3)$$

Where x_l is the input of the l -th layer, x_{l+1} is the input of the $l+1$ -th layer, and $F(x_l, W_l)$ is the residual function of the l -th layer.

3.4 Loss function

In the field of deep learning, the choice of loss function is crucial to model training and performance. Traditional loss functions, such as cross-entropy loss function [31], Focal Loss, and MultiMarginLoss, perform well in many tasks. However, in specific scenarios, they may not fully meet the needs. Therefore, in order to better solve some challenges in deep learning, researchers have been constantly exploring new loss function designs. We propose an innovative approach to combine loss functions, specifically by combining the cross-entropy loss function with an inconsistency loss function [32]. This innovation aims to overcome the shortcomings of traditional loss functions under certain conditions and further improve the performance of the model in certain tasks.

The cross-entropy loss function is a loss function commonly used in classification tasks. It measures the difference between the model's output probability distribution

and the actual labels. For each sample, the cross-entropy loss function takes the logarithm of the probability of the true label, then multiplies and inverses the output probability of the model, and finally sums up to get the loss value. This loss function has good mathematical properties for the use of gradient descent optimization algorithms, prompting the model to better fit the training data, and is especially suitable for multi-category classification problems. However, it may have limitations when dealing with imbalanced data, noise, and special scenarios, which prompts researchers to explore innovative loss function designs to improve model performance. The formula of cross entropy loss function:

$$L = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M y_{ij} \log(y'_{ij}) \quad (4)$$

N is the number of samples, M is the number of categories, y_{ij} is the true label of sample i , and y'_{ij} is the model's predicted probability that sample i belongs to category j .

Inconsistency loss function, which consists of three parts: subspace inconsistency, position inconsistency and representation inconsistency. Subspace inconsistency means that different heads should focus on different subspaces, which is achieved by calculating the subspace similarity between different heads. Position inconsistency means that different heads should focus on different positions, which is achieved by calculating the position overlap between different heads. Representation inconsistency means that different heads should produce different output representations, which is achieved by calculating the output representation similarity between different heads. The inconsistency loss function can effectively improve the diversity of multi-head attention, thereby improving the performance and generalization ability of the model. The formula of subspace inconsistency loss function is:

$$\text{subspace_inconsistency} = \frac{1}{H(H-1)} \sum_{i=1}^H \sum_{j \neq 1}^H \| \text{Proj}(W_i) - \text{Proj}(W_j) \|_F \quad (5)$$

The position inconsistency loss function formula is:

$$\text{position_inconsistency} = \frac{1}{H} \sum_{i=1}^H \| \text{softmax}(W_i Q) - \text{softmax}(W_i K) \|_F \quad (6)$$

The representation inconsistency loss function is:

$$\text{representation_inconsistency} = \frac{1}{H} \sum_{i=1}^H \| \text{softmax}(W_i Q - V) \|_F \quad (7)$$

H is the number of attention heads, $\text{Proj}(W_i)$ represents the projection of the attention weight matrix W_i , Q , K , and V represent the query, key, and value respectively, $\|\cdot\|_F$ is the Frobenius norm to measure differences between matrices.

3.5 Model structure

For telecom fraud text detection, we use Roberta to build word vectors to obtain rich text semantic information, and then multi-head attention focuses on different parts of the input sequence at the same time. Each attention head can focus on capturing different aspects of semantic information. This enables the model to more fully understand

the features and relationships in the text. Then the dropout layer is used to prevent overfitting, and the residual connection layer is used to promote the flow of gradients, accelerate the training convergence of the model, and enhance the stability of the model during the training process, and to a certain extent, help prevent the gradient from disappearing or Explosion problem. Finally, the linear layer classification is performed. **Fig. 1** is the model architecture diagram we designed:

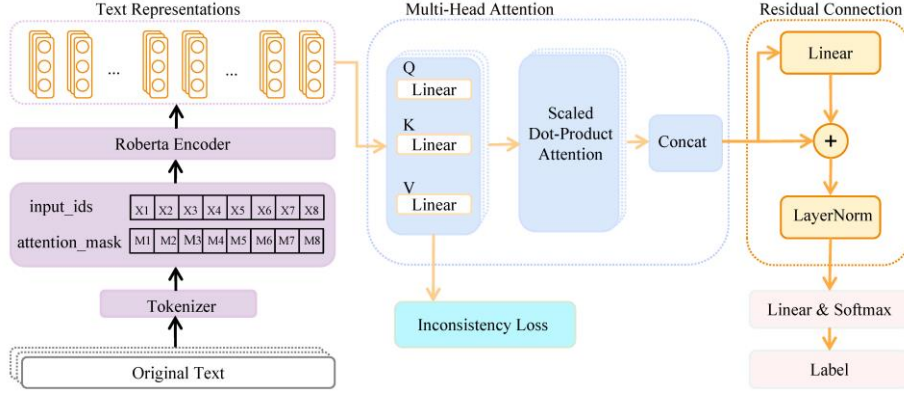


Fig. 1. Structure of model

In the attention layer, we propose Q , K , and V to calculate the inconsistency loss, and then combine them with the cross-entropy loss to form a new loss calculation. **Fig. 2** is the calculation diagram of the loss function we designed.

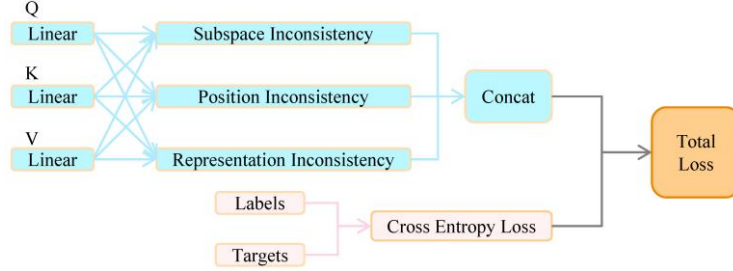


Fig. 2. Structure of loss function

4 Experiments and results

4.1 Dataset

We select 4 types of fraud events that are commonly seen in SMS and phone texts from the CCL2023 telecom network fraud data set [33], such as impersonating customer service, impersonating leadership acquaintances, loans, public security fraud, Familiar people. This data set is a brief description of the case, that is, the victim's transcript,

exported from the anti-fraud big data platform of the public security department. It cannot be directly used for this model detection, because what we need is the text messages or phone calls made by the criminal suspect to the victim. Text information, so we extract keywords from the data, then generate a model and control it by experts, and finally obtain the original text information that is approximately sent by humans. Then we also add some data, which integrates recent fraud events. Our dataset has better real-time performance, such as incorporating related fraud in the epidemic era. We can also add datasets based on recent fraud incidents, such as recent facetime fraud incidents. These 4 types of fraud data are combined with normal text messages and phone calls to form a 5-category data set, with a total of 12,506 pieces of data. The specific category labels and their quantities are shown in **Table 1**.

Table 1. Table of Ours dataset categories and examples

Category	Quantity (unit/item)	Example
normal text	8412	您的外卖订单已成功取消。如您需要帮助或有任何疑问请随时与我们联系。
public security fraud	987	你好，我是法院的工作人员。根据我们的调查取证，发现有人已经起诉了你并提供了你的个人信息和联系方式。现在请你拨打这个电话号码与贷款公司联系清偿欠款事宜。
loans	1001	你好，我是贷款公司的代表。我们提供无抵押、低利率的快速贷款服务。根据你的信用记录和收入情况，我们可以为你办理一笔高额度的贷款业务。
impersonating customer service	1106	您好，我是来自抖音的客服代表。我们注意到您的卡信息可能被误用在了代理商的身份上。为了解决这个问题并保护你的权益，我们需要确认一些细节和进行相应的操作。
impersonating leadership acquaintances	1000	我是你的高中同学，现在急需帮助。我大学同学的妈妈发生了车祸，现在在医院里需要紧急手术费用，但是我的卡因为限额无法直接转账支付医疗费，所以想请你帮忙一下可以吗？

We also use a public data set, FBS data set [8], which is divided into Illegal Promotion class, Advertisement class, others class and Fraud class, with a total of 14058 pieces of data. The specific classification labels and their quantities are shown in **Table 2**.

Illegal Promotion is a subcategory of Advertisement, but it is illegal in China's legal system, so it is split into 4 main categories in total.

4.2 Data processing

When processing our own dataset, we have taken a series of steps to ensure the quality and suitability of the data. Firstly, we segmented the text and counted the frequency of words to better understand the characteristics of the data. Subsequently, we created a stop word list to remove some lower-importance words, such as you, I, he, hello, here, and there. Since our dataset involves telecom fraud, the text information may contain some sensitive information, such as names, ID numbers, phone numbers, bank card numbers, social media accounts, etc. To ensure data security, we anonymized these sensitive information by replacing them with generic identifiers such as "name", "ID", "phone", etc.

Table 2. Table of FBS dataset categories and examples

Category	Quantity (unit/item)	Example
Illegal Promotion	2573	哈啰 DIGIT 号彩票网五周年庆 NAME 您充 DIGIT 元即送倍流水即可取款网址 URL 凭手机号联系在线客服申请
Advertisement	8322	NAME 旗舰店疯狂年中促折后再满减更有免单等你拿聚惠仅限天日 DIGIT 点开抢 URL 退订回 n
Fraud	3065	中国银行尊敬的中国银行信用卡用户我行邀请您提高您的信用卡固定额度请进入 URL 填写申请
others	98	尊敬的用户恭喜您成功注册咪咕帐号您注册的咪咕帐号可以自由登录所有咪咕业务

Additionally, considering that telecom fraud often involves inducing victims to download specific software, we uniformly replaced the software names with the generic term "software". Moreover, we found that some URLs are present in the text. To protect data privacy, we replaced these URLs with 'URL'.

Furthermore, to address potential misspelling issues in fraud-related texts, we employed a Chinese misspelling correction algorithm to help restore the text content. For example, "我的卫星号是" may actually be "我的微信号是".

Lastly, to standardize the length of the dataset, we added or removed some text appropriately to ensure a general length level of the data. These preprocessing steps help ensure that our dataset is clean, secure, and provides a reliable foundation for subsequent analysis and modeling. During the experiment, we divided the dataset by category to ensure that the training and testing sets had the same proportion of samples from each category. We allocated 80% of the data for training and 20% for testing, ensuring that there was no overlap between the testing set and the training or validation sets.

4.3 Model description and parameter settings

Model Architecture For our baseline models, we employ pre-trained transformer-based language models, namely BERT and RoBERTa. Specifically, we use the "hfl/chinese-bert-wwm" and "hfl/chinese-roberta-wwm-ext" variants, which are widely used in natural language processing tasks due to their strong performance on various benchmarks.

Parameter Settings We have extensively experimented with parameter settings to determine the optimal configuration. The batch size is set to 8, with training conducted over 5 epochs and a learning rate of 0.00001. At the model layer, we employed a 12-head attention mechanism, complemented by a dropout rate of 0.3 to prevent overfitting. Furthermore, we explored various combinations of loss functions. Ultimately, we found that augmenting the cross-entropy loss with three inconsistency loss functions, each scaled by a factor of 0.1, yielded the most favorable experimental outcomes.

4.4 Experimental results

This article mainly uses precision, recall and F1 value as evaluation indicators to evaluate the Roberta-MHARC model for detecting Chinese SMS Fake Base Station (FBS) telecom fraud data set and the data set we constructed. In order to performance of the model, we selected the BERT benchmark, Roberta benchmark, logistic regression algorithm proposed by [8], Bert GCN proposed by [6] etc. for comparison. The experimental results are as shown in the **Table 3**.

Table 3. Table of comparison of various model performances

	FBS			Ours		
	Precision	Recall	F1	Precision	Recall	F1
Bert	95.25	94.84	94.97	94.68	94.60	94.56
Bert+Bi-LSTM	96.19	95.94	96.05	95.22	95.12	95.09
Bert+Bi-GRU	95.71	95.62	95.65	95.17	94.92	94.98
Roberta	97.08	96.76	96.88	96.76	96.60	96.57
Roberta+Bi-LSTM	97.14	96.83	94.29	96.96	96.76	96.74
Roberta+Bi-GRU	96.75	96.55	96.64	97.15	97.08	97.03
Logistic Regression	94.90	94.64	94.32	-	-	-
Bert+GCN	96.62	93.56	92.68	-	-	-
Bert-MHARC	96.21	95.91	96.03	95.33	95.24	95.15
Roberta-MHARC	97.90	97.47	97.65	98.11	98.12	98.10

Roberta-MHARC shows strong performance on both datasets. On the FBS data set, Precision reaches 97.90%, Recall reaches 97.47%, and F1 value reaches 97.65%. On our data set, Precision reaches 98.11%, Recall reaches 98.12%, and F1 value reaches 98.10%. This excellent performance is mainly attributed to the Roberta model's use of an extensive pre-training corpus, which enhances word semantic information and can better learn contextual relationships in text. This shows that using large-scale pre-trained models can significantly improve the performance of multi-class classification tasks of Chinese text messages. At the same time, multi-head attention is the main body of our model. Adding multi-head attention mechanism significantly improves the

model performance. The main reason is that the introduction of multi-head attention strengthens the model's modeling ability of the input sequence. Multi-head attention allows the model to focus on different parts of the input simultaneously, and each attention head is able to capture information about specific semantics or locations, thereby improving the model's perception of complex contexts and relationships. By computing multiple attention heads in parallel, the model can more comprehensively understand the characteristics of the input sequence, effectively solving the problems of long-distance dependency and global relationship modeling, thereby performing better on tasks. We also compare the impact of different loss functions on the performance of the Roberta-MHARC model to understand the performance indicators of these models, thereby providing optimization strategies for practical applications. As can be seen from the **Table 4** and **Table 5**. The Roberta-MHARC model achieved the best experimental results on both datasets by using a combination of cross-entropy loss function and inconsistency loss function.

Table 4. Table of experimental results for different loss functions in FBS dataset

	Precision	Recall	F1
MultiMarginLoss	95.94	95.37	95.53
Focal Loss	95.15	94.94	95.01
Combination of cross-entropy and inconsistency	97.90	97.47	97.65

Table 5. Table of experimental results for different loss functions in Ours dataset

	Precision	Recall	F1
MultiMarginLoss	95.18	95.04	95.06
Focal Loss	95.01	95.00	94.93
Combination of cross-entropy and inconsistency	98.11	98.12	98.10

We also conducted ablation experiments to verify the role of each module in the model. We have designed 3 combinations, namely:

Experimental group 1: Roberta+multi-head attention, a multi-head attention mechanism is added to improve the model's modeling ability of different positions and semantic relationships.

Experimental group 2: Roberta+multi-head attention+residual connection. Based on experimental group 1, residual connection is added to promote the flow of gradients and improve training stability. The rationality of the model is further verified.

Experimental group 3: Roberta+multi-head attention+inconsistency loss function. Based on the experimental group 1, the inconsistency loss function is introduced to increase the model's diversity modeling of input data.

The experimental results are shown in the following **Table 6**.

In terms of training speed, our dataset's final model has shown some promising results. Compared to Experiment Group 1, our model's training speed improved by ap-

proximately 10%. Compared to Experiment Group 2, our model's training speed remained steady. Despite the lack of significant improvement, our model still maintained comparable training efficiency to Experiment Group 2. In comparison with Experiment Group 3, our model demonstrated an approximate 4% increase in training speed. This indicates that our model can utilize computational resources more efficiently, accelerating the convergence speed and completing more training steps within the same timeframe. Additionally, on the FBS dataset, our model also exhibited satisfactory training speed advantages. Compared to Experiment Group 1, our model's training speed increased by 3.6%, and by 2% compared to Experiment Group 2. Although it remained on par with Experiment Group 3, we still maintained a certain level of training efficiency. It is worth noting that, despite increasing the model's complexity, our training time did not increase but rather decreased to some extent. This is due to our careful design and adjustment of the model to ensure that performance improvements do not significantly increase training time. This further validates the efficiency.

Table 6. Table of ablation experiment results

	FBS			Ours		
	Precision	Recall	F1	Precision	Recall	F1
Roberta+multi-head attention	97.20	96.94	97.02	97.32	97.38	97.31
Roberta+multi-head attention+residual	97.27	96.97	97.09	97.92	97.97	97.93
Roberta+multi-head attention+inconsistency loss	94.57	95.26	94.29	97.53	97.44	97.42
ResRoberta-MHA (Residual Roberta with Multi-Head Attention)	97.90	97.47	97.65	98.11	98.12	98.10

The ablation experimental results show that the introduction of residual connection and inconsistency loss function plays a certain positive role in improving model performance. Residual connections improve training stability by promoting the flow of gradients; The inconsistent loss function increases the model's diversity in modeling input data.

5 Conclusion

The research focus of this article is on the detection of telecom fraud texts, so the Roberta-MHARC model is proposed and a data set with tens of thousands of items is constructed. First, we demonstrated the strong performance of the model on two data sets, using multi-head attention to increase the model's understanding of text, and using residual connections to promote the flow of gradients and accelerate the model's training convergence. As for the loss function, we use a combination of the inconsistency loss function and the cross-entropy loss function, and experimental results show that this is effective.

In experiments, we observed that the introduction of the multi-head attention mechanism had a significant positive impact on model performance. This makes the model more adaptable to complex task scenarios and effectively improves performance. However, it is worth noting that the effect of the multi-head attention mechanism may vary for different tasks and data sets. Therefore, in practical applications, the specific application scenarios of the model need to be carefully considered to ensure that the introduction of multi-head attention has a positive impact on performance. And in the experiment on the FBS data set, the experimental effect of the multi-head attention mechanism + inconsistency loss function actually declined, but no such situation occurred on the Ours data set, so we believe that in some cases, the inconsistency loss function may introduce task-irrelevant information, affecting the model's performance on the target task.

Although the multi-head attention mechanism achieved very good experimental results, this study also has some limitations. First of all, the data sets and tasks used in the experiment may have specific characteristics. The number of these two data sets only exceeds 10,000, and it is difficult to cover the entire telecom fraud text category. In future research, it should be necessary to construct a broader data set covering more categories and verify the model performance on it. Secondly, although the introduction of the multi-head attention mechanism improves model performance, it also increases the complexity of the model and increases the model training time. In some scenarios, model complexity can become a training and deployment challenge, especially in resource-limited environments. Therefore, further experiments are needed to optimize the complexity of the model.

Finally, this study only focused on the impact of the multi-head attention mechanism, residual connection, and inconsistency loss function, and did not consider other possible improvements. In future work, different model components and structures can be further studied to gain a comprehensive understanding of how to optimize model performance.

Acknowledgments. This work was supported by the Guangxi Natural Science Foundation (No. 2022GXNSFBA035510), Guangxi Key Research and Development Program (No. Guike AB23075178), the National Natural Science Foundation of China (No. 62267002), Innovation Project of GUET Graduate Education (No. 2024YCXS046).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Wu Xiaoli, "White Paper on Preventing and Controlling Telecommunications Network Fraud in the Information and Communications Industry (2023)", China Consumer News·China Consumer Network (2023)
2. Zhang Haijiao, "The public security organs have achieved significant results in cracking down on telecommunications network fraud crimes, and have urgently intercepted 328.8 billion yuan of funds involved in the case", China Economic Net (2024)

3. Jiang Tongtong. "Research and application of NLP-based fraud phone identification method". Shandong Jianzhu University (2022)
4. Rodríguez J F, Papale M, Carminati M, et al. "A natural language processing approach for financial fraud detection". Proceedings of the Italian Conference on Cybersecurity (ITASEC 2022) pp. 135-149 (2022)
5. Zhang G, Li Z, Huang J, et al. "efraudcom: An e-commerce fraud detection system via competitive graph neural networks". ACM Transactions on Information Systems, vol. 40, pp. 1-29 (2022)
6. Zhang X, Huang R, Jin L, et al. "A BERT-GCN-Based Detection Method for FBS Telecom Fraud Chinese SMS Texts". Proceedings of the 2023 4th International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI), pp. 448-453 (2023)
7. Zhou Junjie. "Research on fraud phone text classification method based on deep learning". Shandong Jianzhu University (2023)
8. Zhang Y, Liu B, Lu C, et al. "Lies in the air: Characterizing fake-base-station spam ecosystem in China". Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, pp. 521-534 (2020)
9. Chang J W, Yen N, Hung J C. "Design of a NLP-empowered finance fraud awareness model: the anti-fraud chatbot for fraud detection and fraud classification as an instance". Journal of Ambient Intelligence and Humanized Computing, vol. 13, pp. 4663-4679 (2022)
10. Choi J, Kim J, Lee H. "Hybrid Fraud Detection Model: Detecting Fraudulent Information in the Healthcare Crowdfunding". KSII Transactions on Internet & Information Systems, vol. 16, pp. 1006-1027 (2022)
11. Jurgovsky J, Granitzer M, Ziegler K, et al. "Sequence classification for credit-card fraud detection". Expert Systems with Applications, vol. 100, pp. 234-245 (2018)
12. Xuan S, Liu G, Li Z, et al. "Random forest for credit card fraud detection". 15th International Conference on Networking, Sensing and Control (ICNSC), pp. 1-6 (2018)
13. Ogala, Emmanuel & Akoh, Rose & Agbane, Mohammed & Adeiza, Ezekiel & Tenuche, Bashir & Sunday, Salihu. "Detecting Telecoms Fraud in a Cloud-Based Environment by Analyzing the Content of a Phone Conversation". Asian Journal of Research in Computer Science, vol. 4, pp. 115-131 (2022)
14. Zhao Q, Chen K, Li T, et al. "Detecting telecommunication fraud by understanding the contents of a call". Cybersecurity, vol. 1, pp. 1-12 (2018)
15. Ileberi E, Sun Y, Wang Z. "A machine learning based credit card fraud detection using the GA algorithm for feature selection". Journal of Big Data, vol. 9, pp. 1-17 (2022)
16. Dornadula V N, Geetha S. "Credit card fraud detection using machine learning algorithms". Procedia computer science, vol. 165, pp. 631-641 (2019)
17. Zaremba, W. Sutskever, I. Vinyals, O. "Recurrent neural network regularization". arXiv:1409.2329 (2014)
18. Graves, Alex, and Alex Graves. "Long short-term memory". Supervised Sequence Labelling with Recurrent Neural Networks. Vol. 385, pp. 37-45 (2012)
19. Chung J, Gulcehre C, Cho K H, et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling". arXiv:1412.3555 (2014)
20. Vaswani A, Shazeer N, Parmar N, et al. "Attention is all you need". arXiv:1706.03762 (2017)
21. Yan J. "Intelligent Algorithms for Identification and Defense of Telecommunication Network Fraudulent Call Information under Legal System". International Journal of Network Security, vol. 25, pp. 277-284 (2023)

22. Li X, Yu W, Luwang T, et al. "Transaction fraud detection using gru-centered sandwich-structured model", 2018 IEEE 22nd International Conference on Computer Supported Co-operative Work in Design ((CSCWD)), pp. 467-472 (2018)
23. Meng Z, Xie Y, Sun J. "Detecting Credit Card Fraud by Generative Adversarial Networks and Multi-head Attention Neural Networks". IAENG International Journal of Computer Science, vol. 50, pp. 381-387 (2023)
24. Zhou J, Xu H, Zhang Z, et al. "Using Recurrent Neural Network Structure and Multi-Head Attention with Convolution for Fraudulent Phone Text Recognition". Computer Systems Science & Engineering, vol. 46, pp. 2277-2297 (2023)
25. Devlin J, Chang M W, Lee K, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". arXiv:1810.04805 (2018)
26. Liu Y, Ott M, Goyal N, et al. "Roberta: A robustly optimized bert pretraining approach". arXiv:1907.11692 (2019)
27. Sanh V, Debut L, Chaumond J, et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". arXiv:1910.01108 (2019)
28. Yang X, Zhang C, Sun Y, et al. "FinChain-BERT: A High-Accuracy Automatic Fraud Detection Model Based on NLP Methods for Financial Scenarios". Information, vol. 14, pp. 499 (2023)
29. Jishnu K S, Arthi B. "Phishing URL detection by leveraging RoBERTa for feature extraction and LSTM for classification", 2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), pp. 972-977 (2023)
30. Shafiq M, Gu Z. "Deep residual learning for image recognition: A survey". Applied Sciences, vol. 12, pp. 8972 (2022)
31. Wang Y, Ma X, Chen Z, et al. "Symmetric cross entropy for robust learning with noisy labels", Proceedings of the IEEE/CVF international conference on computer vision. pp. 322-330 (2019)
32. Li J, Tu Z, Yang B, et al. "Multi-head attention with disagreement regularization". arXiv:1810.10183 (2018)
33. Sun C J, Ji J, Shang B, et al. "Overview of CCL23-Eval Task 6: Telecom Network Fraud Case Classification", Proceedings of the 22nd Chinese National Conference on Computational Linguistics, vol. 3, pp. 193-200 (2023)