Cybersecurity

RESEARCH Open Access

Detecting telecommunication fraud by understanding the contents of a call

Qianqian Zhao[1,2*], Kai Chen[1,2*], Tongxin Li[3], Yi Yang[1,2] and XiaoFeng Wang[4]

Abstract

Telecommunication fraud has continuously been causing severe financial loss to telecommunication customers in China for several years. Traditional approaches to detect telecommunication frauds usually rely on constructing a blacklist of fraud telephone numbers. However, attackers can simply evade such detection by changing their numbers, which is very easy to achieve through VoIP (Voice over IP). To solve this problem, we detect telecommunication frauds from the contents of a call instead of simply through the caller's telephone number. Particularly, we collect descriptions of telecommunication fraud from news reports and social media. We use machine learning algorithms to analyze data and to select the high-quality descriptions from the data collected previously to construct datasets. Then we leverage natural language processing to extract features from the textual data. After that, we build rules to identify similar contents within the same call for further telecommunication fraud detection. To achieve online detection of telecommunication frauds, we develop an Android application which can be installed on a customer's smartphone. When an incoming fraud call is answered, the application can dynamically analyze the contents of the call in order to identify frauds. Our results show that we can protect customers effectively.

Keywords: Telecom fraud, Fraud detection, Natural language processing, Machine learning

Introduction

With the development of the Internet, while people are enjoying various kinds of services from the Internet, their private information is gradually leaked out. If a person's privacy is held by attackers, he could be the target of telecommunication frauds. Latest statistics in 2017 show that 90% of smartphone users in China have experienced telecommunication fraud (Facts 2017). According to the data released by the Ministry of Public Security, during the decade between 2006 and 2016, telecommunication fraud cases in China have been growing at a rapid rate of 20% to 30% every year, which have been grown rapidly especially in the past 5 years. According to CNNIC (China Internet Network Information Center), the number of fraudulent calls reported by domestic users in 2015 reached 306 million times, which is 4.25 times that of 2014 (2015 China Mobile Internet Users' Network Security Status Report, China Internet Network Information Center (CNNIC) 2016). At the same time, the telecommunications fraud detection became a hot topic. Therefore, the Chinese government has built related departments and financed to help the research on telecommunication fraud detection (Li and Yuan 2017). Since the telecommunication frauds cause severe financial loss to telecommunication customers, it is necessary and urgent to detect telecommunication frauds.

In order to detect telecommunication frauds, most of the current approaches are based on labeling the caller numbers that are identified as frauds by customers. At the same time, there are also many researchers who use machine learning techniques to detect fraudulent calls. They select features based on factors such as phone numbers and call types. They use machine learning algorithms to train models, and use these models to detect fraudulent calls, which can also achieve good detection accuracy. However, as the number change software is widely used, fraudsters use software to change their phone number constantly or disguise their number as the official number of government agencies. These reasons make it possible for conventional telephone number-based detection methods can be easily bypassed.

* Correspondence: qian01fly@163.com; chenkai@iie.ac.cn
[1]SKLOIS, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
Full list of author information is available at the end of the article

In this paper, we put forward an approach to detect telecommunication fraud through analyzing the contents of a call. However, this is quite challenging, mainly due to the complexity of the contents of a call impedes the analysis. To solve this problem, in a nutshell, we learn the telecommunication fraud from the reports and news on the Internet for understanding the contents of a call. Particularly, first, we collect descriptions of telecommunication fraud from the Internet. In our study, we gather 12,368 samples of telecommunication frauds from Sina Weibo (2017) and 3234 samples from Baidu (2017). Then, we filter out those that have no contents of the fraud calls. Leveraging machine learning algorithms, we analyze textual data that related to telecommunication fraud and train models to select textual data. The prediction accuracy of the machine learning model is 98.53% on the dataset we selected. In the next step, enlightening by decision tree algorithm, we use Natural Language Processing (NLP) techniques to extract features from the data. After extracting features, we build detection regulations to prove whether a piece of text is related to telecommunication fraud or not. Finally, we develop an Android application for the purpose of telecommunication fraud early warning using the generated detection rules. In this way, we are able to detect telecommunication fraud through the contents of a call instead of relying on the blacklist. At the same time, we do not need to upload any information of the user to a remote server. All the operations are handled locally.

Contribution

The contributions of the paper are summarized as follows:

- A new technique for telecommunication fraud detection. Instead of relying on constructing a blacklist of fraud numbers, we identify telecommunication frauds only through the contents of a call. Features of previously reported telecommunication frauds are extracted using natural language processing techniques, which serves to detect further telecommunication frauds.
- We implement an Android app to perform online detection of telecommunication frauds. This app uses speech recognition techniques to identify the content of the call and justifies fraud calls based on the features before warning the user. We will release the app for protecting telecommunication customers.

Roadmap

The rest of the paper is organized as follows: Section "Related work" gives the related work to our studies. Section "Overview" provides the overview of our approach. Section "Data preparation" illustrates data collection and analysis. Section "Feature extraction and rule building" describes how to extract features and builds rules for telecommunication fraud detection. Section "Application implementation" illustrates the implementation of our Android application to detect telecommunication frauds. Section "Evaluation" illustrates the results of our experiments. Section "Discussion and Conclusion" discuss and summarize our approach, respectively.

Related work

Telecommunication fraud detection

Recently, telecommunications fraud detection has become a hot research direction gradually. Some researchers have used blacklisting and whitelisting methods to prevent telecommunications fraud (Jiang et al. 2012; Zhang & Fischer-Hubner 2011; Patankar et al. 2008; Wang et al. 2007). More researchers use machine learning techniques to determine if they are malicious. They extract a variety of features for malicious call detection, and most of the features include telephone numbers, call-time, domain names, call networks, and the actions of listeners and callers, etc. (Kolan et al. 2008; Azad & Morla 2011; Azad & Morla 2013; Jiang et al 2013; Leontjeva et al. 2013; Rebahi & Sisalem 2005; Rebahi et al. 2006; Sorge & Seedorf 2009; Srivastava & Schulzrinne 2004; Wang et al. 2013; Wu et al. 2009; Zhang and Gurtov 2009). In 2015, Subudhi and Panigrahi (2015) published their research on telecommunication fraud using the features of a telephony communication as the input and Quarter-Sphere Support Vector Machine to distinguish fraudulent calls. The input features include call duration, call type, call frequency, location and time, and it has achieved good recognition accuracy. Then, in 2017, they used a type of C-means clustering for telecommunication fraud detection again (Subudhi and Panigrahi 2017), and got a good result as well. Coincidently, Li et al. (2018) published an article on telecommunication fraud detection in recently as well. They used machine learning algorithm to detect malicious calls, and they extracted features of calls just similar to S. Subudhi et al. It seems difficult for these approaches to detect fraudulent calls with unlabeled new numbers. On the other hand, the researches on the content of conversations are rare. The work of Miramirkhani et al. (2017) is representative, and they conducted the research on the technical support scams. They recorded and analyzed the voice content of the phishing telephone fraud and malicious webpages which triggered such scams. Though they eventually put forward simple functions to help users keep away from malicious web pages, their approach did not help avoiding malicious calls directly.

Nowadays, more and more fraud criminals use change number software to constantly change their phone numbers in China. At the same time, there are also many fraudsters who use number-changing software to disguise their numbers as government agencies' numbers, such as procuratorates and police stations. Therefore, traditional feature detection based on phone numbers is no longer reliable and can easily be circumvented by fraud. Therefore, we have built a content-based telecommunication fraud detection method. Our research is content-based approach without relying on incoming caller numbers and related information. Therefore, it can detect fraudulent calls with any phone number. Furthermore, the majority of former studies are conducted on the assumption which the telecommunication network may provide more information, while our research relies on the clients entirely, where we only needs the content of the call to determine if it is a fraudulent call.

Fraud detection

Fraud detection has always been the subject of some surveys and commentary articles because of the severe damage to the society. Delamaire et al. (2009) proposed different types of credit card frauds, such as bankruptcy fraud, theft fraud/counterfeit fraud, application fraud and behavioral fraud, discussing the feasibility of various techniques to combat this type of fraud, such as decision tree, genetic algorithms, clustering techniques and neural networks. Rebahi et al. (2011) proposed the VoIP fraud and the fraud detection systems to it checking their availability in VoIP environments in various fields. These detection systems are classified as two categories: rule-based supervised and unsupervised methods. Lookman Sithic and Balasubramanian (2013) investigated the categories of fraud in medical field and vehicle insurance systems. Various types of data mining techniques were used to detect fraud in these areas according to the results. The financial fraud detection has become the most popular topic in the area of fraud detection (Abdallah et al. 2016) which usually leads to high economic losses.

Overview

In this section, we present an overview of the telecommunication fraud problem and our solution, and we will explain our idea of telecommunication fraud detection briefly. This article's purpose to the detection of telecommunication fraud is that they could be warned by the notification from the application on Android platform when users receive fraudulent calls. The whole process is divided into three parts: the first is the collection and pre-processor of telecommunication fraud data. The second part is extracting features and building rules of detection. The last part is the implementation of telecommunication fraud alert applications. The overview of our approach is shown in Fig. 1.

The first step is the collection of telecommunications fraud data. In order to analyze the characteristics and modes of telecommunication fraud, the first thing to do is collecting textual data. Data collection is mainly to collect telecommunication fraud-related texts. The target data includes the case of fraudulent calls, the description language of telecommunication fraud, and the news on the media. In the data collection process, web crawler technology is used to collect data, and search engines (such as Baidu, etc.) are helped to collect textual data on telecommunication fraud on the Internet.

The second step is feature extraction and rule-building. After the data collected in the first step, it is important to extract the features and build rules for detecting telecommunication fraud. This research uses natural language processing technology to extract features which are keywords from fraud text. And we use machine learning algorithms to prove the appropriateness of textual data we collected and the validity of keywords we extract. Then, according to the features which are extracted from the text, this research builds the detection rules of telecommunication fraud.

The last part is the implementation of telecommunication fraud detection. In this paper, we developed a telecommunication fraud alert application on the Android platform. In detail, the application first starts to monitor the incoming call when a call coming to the users' phone. Then the application uses speech recognition technology to convert the caller's voice into text. After that, the application uses the detection rules that built in the previous step to determine if it is a fraudulent call or not. If the application predicates that it is a fraudulent call, a warning information will pop up on the smartphone's screen to prompt the user to pay attention to this call.

Data preparation

This section is mainly described the process of how we collect textual data and how we selected data by machine learning algorithms. This section divided into two parts. The first part is data collection and preprocessing and the second part is how we select data by machine learning algorithms. In the first part, we describe the data source, data collection, and data preprocessing. In the second part, we analyze textual data that collect from two data sources by machine learning algorithms and choose one data set of two to continue next research.

Data collection and preprocessing

The quality of the data determines the result of the detection. Therefore, it is important to construct dataset. After finishing investigation and research, we choose Sina Weibo and Baidu as the source for collecting telecommunication fraud data. The reasons are as follows. First, Sina Weibo is the largest open communication platform in China. With a wide range of data covering various areas, and everyone has access to these data; Second, Baidu is the largest search engine in China, and its powerful search

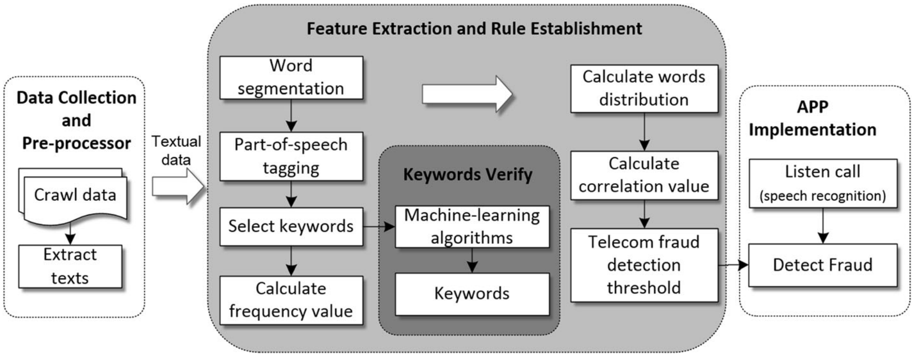

Data Collection and Pre-processor
Crawl data
Extract texts
Textual data
Feature Extraction and Rule Establishment
Word segmentation
Part-of-speech tagging
Select keywords
Calculate frequency value
Keywords Verify
Machine-learning algorithms
Keywords
Calculate words distribution
Calculate correlation value
Telecom fraud detection threshold
APP Implementation
Listen call
(speech recognition)
Detect Fraud

Fig. 1 Overview

capabilities in Chinese can bring us high-quality data with a big amount.

We build crawlers to collect data information from Sina Weibo and Baidu search including textual information such as a description of the fraudulent call that the user has received and related news reports. Finally, we collected 12,368 samples of text from Sina Weibo and 3234 samples of text from Baidu.

After the data collection is completed, the textual data is preprocessed to do the further analysis. Data preprocessing is mainly divided into two parts. The first part is formatting, the main work is to remove meaningless characters and duplicate values. We use the method of measuring text similarity to remove duplicate values. When the similarity between two pieces of text reached 80%, one of them would be treated as the duplicate value. The second part is manual filtering, and the main task is to manually filter data that meet the subject of telecommunications fraud. After data preprocessing, we select 647 fraud-related textual data from Sina Weibo and 1443 textual data from Baidu.

Analyzing and selecting data by machine learning algorithms

The purpose of using machine learning to analyze textual data is to lay the groundwork for constructing our own detection methods. The unique advantages of machine learning algorithms in pattern recognition make it a popular choice for analyzing textual data. There are two reasons for choosing machine learning algorithms to analyze data. First, the use of machine learning algorithms to analyze and verify the data first fully proves the feasibility of subsequent ideas. And the idea of telecommunication fraud detection is inspired by machine learning algorithms. We determine whether a call is a fraudulent or not through features. Second, after data collection, it is necessary to verify whether the data we collected fully describe the characteristics of telecommunication fraud and distinguish it from ordinary text. Machine learning could verify the quality of the dataset and help select appropriate dataset.

To convert text into the inputs that machine learning algorithms accept, the textual data would be vectorized first. Firstly, we use the TF-IDF algorithm (Salton and Buckley 1988; Oren 2002) to extract the keywords from all data. The conversion rule is to make each keyword represent a feature, then we segment the text and remove the stop words. If one text contains the keyword which represented by this feature, the value of this feature would be set to 1, otherwise, it would be set to 0. According to this rule, these texts are converted to n-dimensional vectors.

On the next stage, we use these vectors to train the model with machine learning algorithms. We use the n-dimensional vector and data label Y as inputs so that machine learning algorithms output the models. Then we use these models to predict new data and test the accuracy of these models. Meanwhile, we use the 7-fold and 10-fold cross-validation in the model training process. The algorithms used in this study include logistic regression (Preacher et al. 2006), neural network (Schwenk and Gauvain 2005), and decision tree (Blockeel et al. 2006a). The results of data analysis are shown in section “Evaluation”.

Feature extraction and rule building

This section contains two parts, they are feature extraction part and rules building part. In the feature extraction part, we describe how we extract features from the textual data. In the rule building part, we explain how we build rules based on the extracted features. This section fully expresses our ideas on telecommunication fraud detection which is the core content of this paper.

Feature extraction

In other researchers’ methods of telecommunication fraud detection, the features they extracted are mainly calling numbers, calling times, calling types, etc. Our research is based on the identification of call content to detect fraudulent calls. Hence, we need to extract features from more complex data, and this is a challenge to the extraction method. For text-type information, Natural Language Processing (Jackson and Moulinier 2007) is a suitable technique. The techniques of word segmentation and part-of-speech-tagging in Natural Language Processing in the Chinese domain can handle Chinese content well. It is what we need to deal with the Chinese textual information, so the Natural Language Processing techniques are our preferred choice.

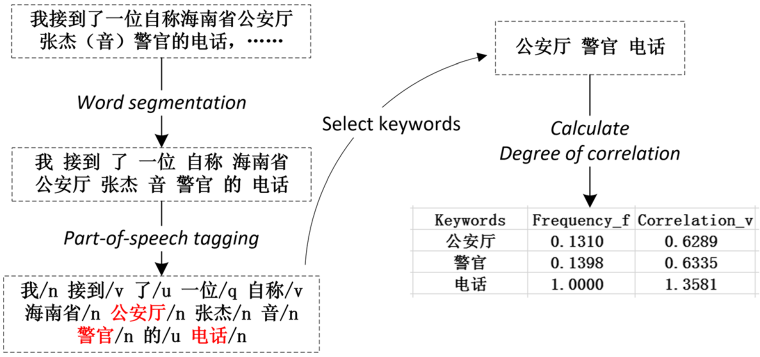
Inspired by the decision tree algorithm in the data analysis phase, we designed a method to extract features from telecommunication fraud-related text by using the Natural Language Processing techniques. The extraction process is shown in Fig. 2.

The first step in feature extraction is to segment the text (Gao et al. 2003). This is because words and words are put together without blank spaces in Chinese, unlike English, where there are spaces between words. Therefore, the first step in Chinese Natural Language Processing is to separate words within sentences. For example, the process divides the sentence "今天天气真好啊!" (The weather today is really good!) into "今天 天气 真 好 啊" (The / weather / today / is / really / good!). Next part is part-of-speech tagging. As its name suggests, tagging attributes are marking attributes to a word, such as nouns, verbs, adjectives, and adverbs.

The next part after that is keywords selection. The process includes various steps: The first step is to remove the stop words (Chen and Chen 2001). Stop words are meaningless words that occur after the word segmentation, such as prepositions just like "on", "to", "of", etc. There are various types of stop words lists on the Internet, and it doesn't exist a special stop words list that can be applied to all natural language studies. Therefore, we establish a new stop words list. Our stop words list includes 1601 words which cover the most common stop words, as well as the words we selected based on the word what is commonly used on the phone, such as Hello, Good, Hang Up, Hold on, and so on. After removing the stop word, we write programs to select keywords based on the part-of-speech. For example, we remove prepositions, adverbs and other meaningless words, at the same time we retain nouns, verbs, and other meaningful words. We continue to filter the keyword list manually after selecting keywords by programs. This step is mainly to remove the words which meaningless such as personal names and geographic names. After that, we get a keywords list extracted from textual data came from telecommunication fraud related data.

After getting the keywords list, we calculate the value of the frequency of keywords in telecommunication fraud data and normal data which are not irrelevant to telecommunication fraud. The normal data comes from the Chinese text classification dataset THUCNews provided by the Tsinghua NLP Group (Sun et al. 2016). THUCNews is generated based on the historical data of the Sina News RSS subscription channel from 2005 to 2011. We selected some of these subsets to our dataset. Then we calculate a fraud tendency value of each keyword called "Degree of correlation" according to the value of the frequency of telecommunication fraud data and normal data. The keywords' distribution of the value of frequency with correlation value is shown in Fig. 3. This is the result of feature extraction.

Figure 3a shows the relationship of keywords' frequency of fraud data and the correlation value representing the data tends to distribute linearly. Note that the correlation between the correlation value and the keywords' frequency of fraud data is high. Figure 3b shows the relationship of keywords' frequency of normal data and the correlation value. On the contrary, this distribution is relatively scattered and it can be seen that the correlation between the correlation value and keywords' frequency of normal data is not high. Figure 3c demonstrates the difference between Fig. 3a and b. From Fig. 3c, we find that these two distributions are clearly different, they have the same amount of data on the Y-axis. The data which has high correlation value has a low value on the frequency of normal data but the high value on the frequency of fraud data. These figures describe the different distributions


我接到了一位自称海南省公安厅
张杰（音）警官的电话，……
Word segmentation
我 接到 了 一位 自称 海南省
公安厅 张杰 音 警官 的 电话
Part-of-speech tagging
我/n 接到/v 了/u 一位/q 自称/v
海南省/n 公安厅/n 张杰/n 音/n
警官/n 的/u 电话/n
Select keywords
公安厅 警官 电话
Calculate
Degree of correlation

Keywords	Frequency_f	Correlation_v
公安厅	0. 1310	0. 6289
警官	0. 1398	0. 6335
电话	1. 0000	1. 3581

Fig. 2 Example of extraction features processing

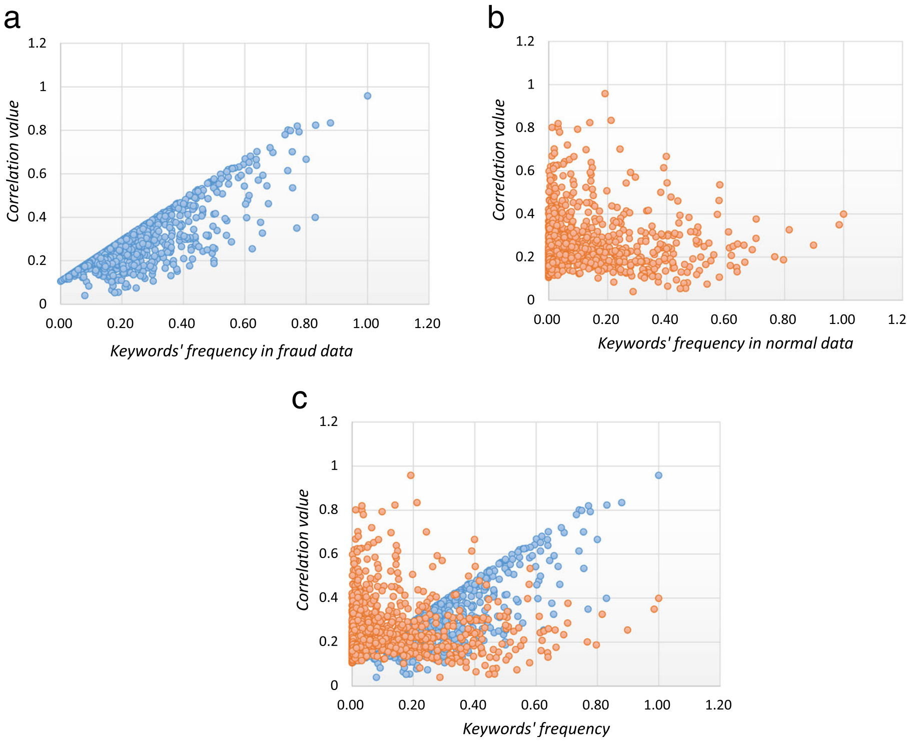


Fig. 3 Distribution of keywords' frequency with the degree of correlation. **a** Distirbution of keywords' frequency in fraud data. **b** Distribution of keywords' frequency in normal data. **c** Distribution of keywords' frequency in fraud data and normal data

of keywords in fraud data and normal data by frequency and correlation values, which proves that the features we extracted before could represent the characteristics of telecom fraud effectively. It also proves that our method of extracting features is appropriate to these textual data.

Detection rules building

After extracting the features, this paper builds detection rules on the basis of features extracted before. This paper proposes a method to detect telecommunication fraud based on the keywords we selected. In the previous data analysis, this study used the words as the basic unit to vectorize the text and use it as the input to the machine learning algorithm. In the same way, this study also uses the keywords as the basic elements in the step of feature extraction. At the same time, this study calculated the frequency of keywords and the correlation value of telecommunication fraud. The next step is how to use these characteristics and values to builds rules for the detection of telecommunications fraud.

In the decision tree algorithm, the features which have a greater impact on the results selected by calculating the information gain (IG) (Blockeel et al. 2006b). In the same way, the keywords which have a high influence on the correlation between telecommunication fraud data and normal data would be selected by calculating the distribution of keywords. Corresponding to the information gain, this study shows that this value is correlation value. And we calculate the correlation value by the difference between keywords' frequency in telecommunication fraud data and normal data.

After that, we build rules for detecting telecommunication fraud. The decision tree algorithm builds nodes and branches by features. Different from the decision tree algorithm, this study sums up the correlation values of the keywords appearing in the text when predicting whether a text is related to telecommunication fraud. When the sum exceeds a threshold, the text is deemed to be related to telecommunication fraud. The formula for detecting rules is as follows:

$$\mathrm{R} = \sum_{i=1}^{w_i \in L} \left[F_f(w_i) - F_n(w_i) \right] - T_k$$

Among them, i represents the subscript of the keyword, w_i represents the keyword detected from testing text, $F_f(w_i)$ and $F_n(w_i)$ are the frequencies of the keyword (w_i) in the fraud data and normal data, L is the keywords' list used for fraud detection, and T_k is the threshold of whether the text is the telecommunication fraud data. When the result R is greater than or equal to 0, the text is estimated as telecommunication fraud data, otherwise, it is a non-fraud data.

Application implementation

After building rules for telecommunication fraud detection, we develop an application on Android platform. The purpose of this application is to warn the user when user receiving a fraudulent call. In other words, we determine whether a call is a fraudulent call by understanding the user's call content and if so, we will alert the user by the application.

Considering most of the smartphone users in China are using Android operating systems, we developed the telecommunication fraud alert application for the Android platform. The workflow of the application is described below. First, users must press a button to start the service which could turn on the service of monitoring incoming calls. Second, when the user receives a call and the call connected, the application starts to record the voice on the phone. Next, the application transfers the audio files to the module of speech recognition provided by iFLYTEK (Open platform from iFLYTEK 2017). Then the speech recognition module will return to the textual result. In the next step, the fraud detection module will detect if the call is related to telecommunications fraud by applying the detection rules we built to the text. Once the fraud detection module determines that the call is a fraudulent call, the application would send an alert to user's screen of the smartphone so that it could prompt the risks of fraud to the user. The structure of telecommunication fraud alert application is shown in Fig. 4.

The main interface of the telecommunication fraud alert application is shown in Fig. 5. After the application started, the user clicks the button on the main page. A service will be triggered in the background while a log information said "Service has started" shown on the screen. When the user receives a fraudulent call, the application will go to the foreground and show an alert page. The user could click the "Read more" button to look for the details of the fraudulent calls.

There is a reason for choosing a new approach put in the application we developed instead of machine learning algorithms. There are some differences between the conversations in the real call environment and the descriptive text in our training set. However, machine learning algorithm requires both the training set and the test set have the same distributed consistency pattern. In order to detect telecommunication fraud in the real world on Android platform effectively, we propose a new method instead of machine learning programs.

Evaluation

This section displays the results of our experiments. We use crawler written in Python to collect data from Sina Weibo and Baidu. We collected 12,368 samples of text from Sina Weibo and 3234 samples of text from Baidu. Then we used machine learning algorithms to analyze the textual data. We use decision trees, neural networks, and decision tree algorithms for model training and evaluation. The accuracies of the models are generally above 98.53%. Then we extract features and build rules to detect telecommunication, and we apply detection rules to the application developed on the Android platform. Finally, the application's recognition accuracy could reach 90% to fraudulent calls we tested.

Results of data analysis

This study use crawler programs written in Python to collect data from Sina Weibo (2017) and Baidu (2017). We collected 12,368 samples of text from Sina Weibo and 3234 samples of text from Baidu. And after data preprocessing, we select 647 samples of fraud-related textual data from Sina Weibo and 1443 samples of textual data from Baidu. Then we use machine learning algorithms to analyze the textual data. We use decision trees, neural networks, and decision tree algorithms for model training and evaluation. The numbers of datasets of Sina Weibo and Baidu is shown in Table 1(a) and (b). The accuracy of machine learning models in Table 2(a) and (b).

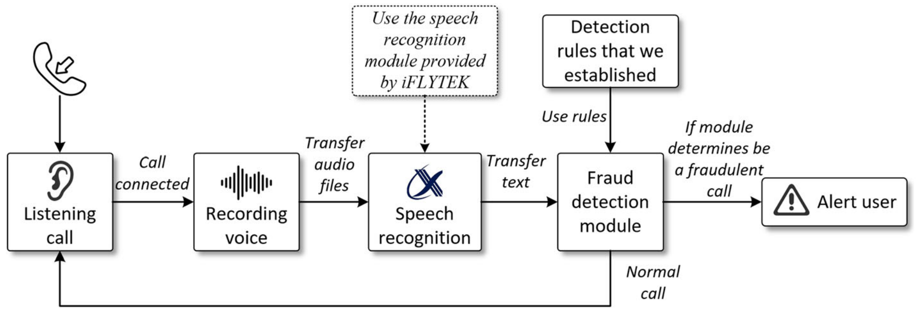
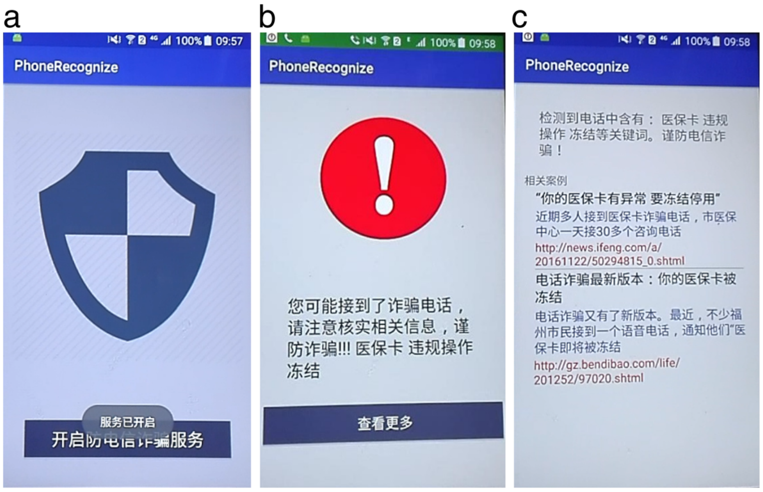


Fig. 4 Structure of telecommunication fraud alert application



您可能接到了诈骗电话，请注意核实相关信息，谨防诈骗!!! 医保卡 违规操作 冻结

检测到电话中含有： 医保卡 违规 操作 冻结等关键词。谨防电信诈骗！

“你的医保卡有异常 要冻结停用”

近期多人接到医保卡诈骗电话，市医保中心一天接30多个咨询电话

http://news.ifeng.com/a/20161122/50294815_0.shtml

电话诈骗最新版本：你的医保卡被冻结

电话诈骗又有了新版本。最近，不少福州市民接到一个语音电话，通知他们“医保卡即将被冻结

http://gz.bendibao.com/life/201252/97020.shtml

Fig. 5 The UI of telecommunication fraud alert application. (**a**)shows the screen after the service is started in the home page. (**b**) shows the performance when the application detects a fraudulent call. (**c**) shows detailed information about that type of fraudulent call

Comparing the forecasting results of Sina Weibo’s data to Baidu data, it can be seen that the quality of the data is important to the training results. The accuracy of all these models for Baidu data exceeds 98.53%, while there are only 80–83% for Sina Weibo data.

There are some reasons result in that the accuracy of Baidu data is higher than Sina Weibo data. First, the single text of Baidu data is much longer than the one in Sina Weibo which means the data from Baidu can express its meaning more comprehensive. Second, amount of data from Baidu is bigger than that from Sina Weibo, which means a better machine learning model would be trained. And the result of data analysis also verifies the assertion that “Data quality determines the quality of training results.” Therefore, we would use the datasets from Baidu for feature extraction.

On the other hand, the features we use in these machine learning algorithms are the keywords extracted from textual data. The high accuracy of prediction also verifies the validity of these keywords and lays the foundation for our process of rule building.

Table 1 Numbers of datasets

Numbers of datasets	Training set	Test set
(a) Numbers of datasets for Sina Weibo		
Telecommunication fraud text	532	115
Normal text	299	185
Total	831	300
(b) Numbers of datasets for Sina Weibo		
Telecommunication fraud text	1100	343
Normal text	2200	541
Total	3300	884

Features analysis

All the features extracted are from the keywords in the text, which means the difference between features is slightly small. The key is to select the keywords which could better represent the telecom fraud. Thus, we calculated the “Degree of correlation” value of each keyword to indicate whether this keyword could better represent the telecommunications fraud. We built a list of keywords ordered the value and each keyword represents a feature. In actual tests, the threshold needs to be set according to the number of keywords. Therefore, we have chosen different numbers of keywords to conduct the experiments instead of selecting all the keywords. After our experiments, we found that as long as the selected keywords reach a certain number, selecting more keywords does not significantly improve the accuracy of the prediction. The difference in the detection accuracy of the rules constructed by different numbers of keywords is within 3%.

Table 2 Analysis results

Model	Accuracy	Precision	Recall	F1
(a) Result of Sina Weibo data analysis				
NN	0.8367	0.7260	0.9217	0.8122
C5.0	0.8033	0.6818	0.9130	0.7806
C5.0(2)	0.8133	0.6879	0.9391	0.7941
CRT	0.8333	0.7407	0.8696	0.8000
(b) Result of Baidu data analysis				
NN	0.9853	0.9797	0.9825	0.9811
NN(2)	0.9921	0.9855	0.9942	0.9898
C5.0	0.9910	0.9912	0.9854	0.9883
C5.0(2)	0.9887	0.9883	0.9825	0.9854

At the same time, we found that the accuracy of the keywords located in the front of the keyword table is better than the accuracy of the keywords in the bottom. And, the keyword which has a smaller correlation value tend to be less important. Therefore, when we select a certain number of keywords for prediction, we select them from the front of the keyword list.

Detection performance using our rules

For the rules proposed in this study, we selected the datasets to test it. First, test the effect of different thresholds on the accuracy of the prediction. As follows, we have selected 200 telecommunications fraud-related data and 200 normal data. These data do not overlap with the data we extracted features from. We set different thresholds to test the effectiveness to the accuracy of the rules. The result is shown in Fig. 4:

In the detection rules shown in Fig. 6a, it can be seen that when the threshold value is set from 17 to 18, the prediction accuracy rate reaches the maximum value. In the experiment shown in Fig. 6a, the keywords to be detected is top 200 selected from keywords list. In the experiment shown in Fig. 6b, the keywords to be detected is top 300 selected from keywords list. In Fig. 6b, when the threshold is 22, the prediction accuracy reaches the maximum value. It can be seen that the difference in the number of selected keywords also causes the difference of threshold value which the best accuracy required.

In order to better illustrate the detection effect of our detection rules, we plot the ROC curve for our rules in Fig. 7. In Fig. 7, we can see that the area under the curve is greater than 0.9 which means that the AUC score is greater than 0.9. It is clear that getting the appropriate threshold value could help the recall rate of the model come up to 90% (y-axis on the ROC curve), which shows that only few of normal cases being predicted to be malicious. Therefore, the ROC curve also proves that our detection rules are effective.

In addition, we select the top 200 keywords and set a threshold of 18, then we select four different test sets to further test. Each test set contained 100 data related to telecommunication fraud and 100 data not related to telecommunication fraud. There is no coincidence between each dataset. All the test data in this part that related to telecommunication fraud come from Baidu. The test results are shown in Table 3.

From Table 3, the accuracy of our detection rules exceeds 93% on most datasets. We can see that the rules we have built are effective. And the recall of results is higher than precision on all four datasets. It means that our detection rules are suitable for use in practice. On the next stage, this study will apply the detection rules to applications on the Android platform. Although not exceeding the accuracy of machine learning algorithms, it has many advantages of its own. The advantages are as follows: (1) The prediction effect is good; (2) The detection procedure is lightweight, which could reduce the system overhead; (3) The detection rules are flexible and easy to update; (4) The calculation can be completed locally to avoid exposing the user's privacy.

Performance of our application

In order to test the performance of telecommunications fraud alert applications, we designed some experiments to detect it. This research designed some telecommunication-fraud-related dialogues, and let the experimental participants read these conversations on the phone. We designed 15 dialogues based on the recordings on the Internet and the descriptions of the parties who received the fraudulent calls. We test 10 samples of the conversations used Mandarin, and 5 conversations used dialects. Among the calls read by mandarin, 9 calls were detected as

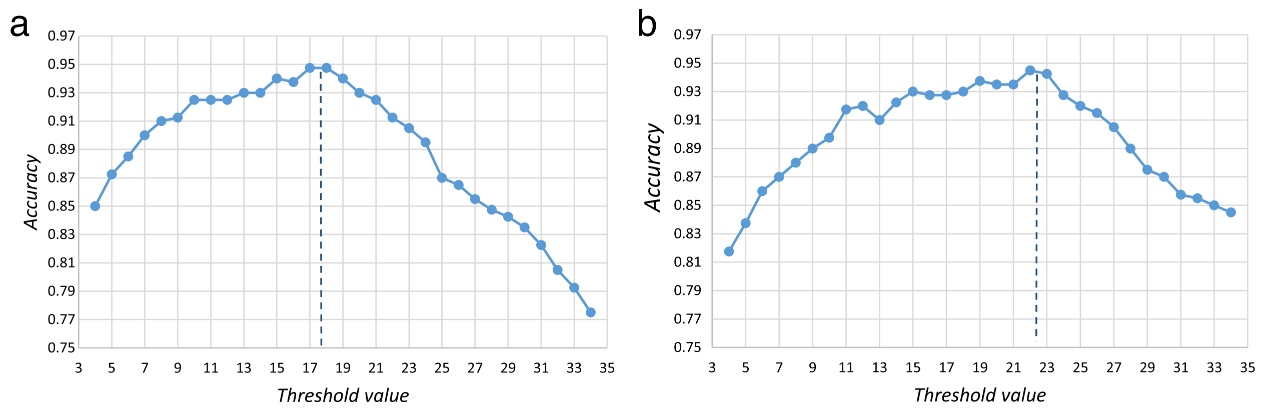


Fig. 6 Line chart of prediction accuracy changes and threshold. **a** Prediction accuracy changes and threshold with 200 keywords. **b** Prediction accuracy changes and threshold with 300 keywords

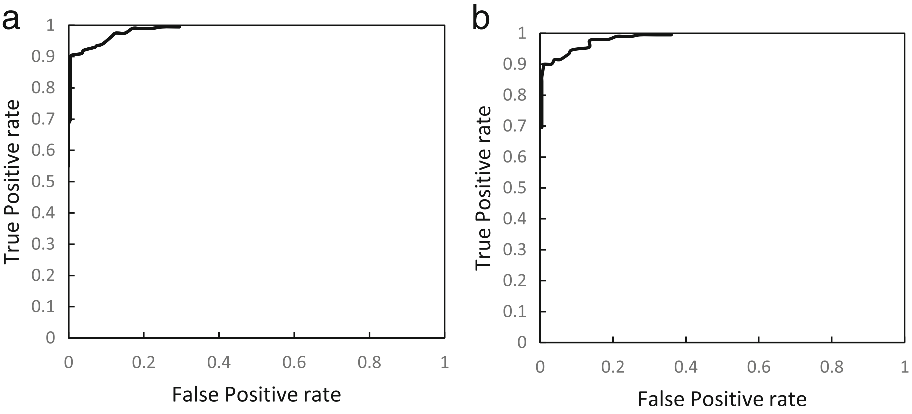


Fig. 7 ROC curve with different number of keywords. **a** Receiver operating characteristic curve with 200 keywords. **b** Receiver operating characteristic curve with 300 keywords

fraud. It indicates that the application has detected 90% of fraudulent calls using Mandarin. Among the calls read by dialect, only 2 calls were detected as fraud. The main reason for this phenomenon is that the detection accuracy depends on the quality of speech recognition. Once the caller uses a dialect with a strong accent, the speech recognition program cannot return to the correct text. In the future, with the development of voice recognition technology and the improvement of recognition accuracy, the detection accuracy of fraudulent calls using dialects by our application will increase.

Discussion

In summary, we built rules for detecting telecommunication fraud by extracting the features. Then, we developed a telecommunication fraud alert application on the Android platform with the rules. And it solved the problem of early warning of fraudulent telecommunications. During the procedures, we evaluated the quality of the data by data analysis. Then we extracted features of words from textual data and built rules for telecommunication fraud detection. After that, our experimental results validate the feasibility of the method. We applied the detection rules to telecommunication fraud alert application so that the application could detect fraudulent calls effectively.

Here we will discuss privacy issues. Since our research is based on the detection of telecommunication fraud by the content of a call, it is necessary for us to obtain the user's call content. However, there are several points would guarantee that we will not disclose the privacy of users. First, the speech recognition technology that we used is not connected to the network. The entire process of speech reorganization can be completed locally. The recordings of users' call are stored in local storage, and the recordings would never be uploaded to the server or network. Second, these call records will be deleted immediately once these calls are judged to be non-fraud. At the beginning of the detection, the user's call recording will be placed in the application's private directory to prevent it from being acquired by other applications. When the detection is completed and the call is not judged to be a fraudulent call, the application would immediately delete the recording file. When detecting that a call is a fraudulent call, the application would ask the user whether to keep the recording file and keeps or deletes the recorded file as the user wishes. These measures ensure the security of user data in the local storage space. In the future, we plan to encrypt the content of the call that the user chooses to save locally to increase the security of the user information. Third, we allow the user to set up a whitelist. For example, calls from the number which comes from the address book will not be recorded. Last but not least, our application would still update the rules periodically to detect the latest fraud cases. While our app doesn't have internet access, we update the detection rules by updating the entire app directly. we can publish the app to the app store, and users use the app store to update apps regularly. We could update the rules to ensure that users are alerted to the recent fraud cases by updating the entire app regularly. That means the entire running process of application does not be connected to the Internet. This ensures that the application will not upload any user information to the Internet.

Our application does not bring dramatically performance degradation to the user's mobile phone. We tested

Table 3 Test results for telecommunication fraud detection rules

Test datasets	Accuracy	Precision	Recall	F1
Dataset 1	0.9300	0.8839	0.9900	0.9339
Dataset 2	0.9800	0.9800	0.9800	0.9800
Dataset 3	0.8500	0.8431	0.8600	0.8515
Dataset 4	0.9500	0.9166	0.9900	0.9519

the application on the real machine. We tested the application on a model phone of Honor 7 from HUAWEI company. The phone's RAM is 3G. Our application occupies about 100 M to 125 M of memory when it is running. The percent of memory space occupied is about 3.35–4.06%. And the application consumes approximately 0.02% of power for half an hour. It can be seen that our application has little influence on the performance of mobile phones. There are two reasons why our applications have better performance. First, our application carries on voice recognition and fraud detection only after having a phone call and connecting successfully. At other times, it only run in the background and seldom takes up the computing resources of the mobile phone. Second, our detection rules are very simple, after finishing the speech recognition, the application just needs to count keywords from text and continues a simple calculation before it concludes. Our method is much simpler than the general machine learning algorithm. Therefore, our application would not affect the performance of the user's mobile phone dramatically.

However, there are some limitations in our research. On the one hand, the amount of data obtained is not large enough, and we did not get enough original recordings of genuine fraudulent calls. Moreover, our speech recognition module is based on the local speech recognition library of the IFLYTEK to recognize the voice, and the recognition accuracy of the local library is lower than that based on the cloud. This also has influences on the detection performance of our application. Also, our application has not been applied into practice yet, it is not sure that if it would get a good performance in practical life.

Therefore, there are still many tasks that can be improved. On the one hand, it would be very helpful for us to obtain more recordings of the real fraudulent calls in the further research. This will help us extract more precise features and improve the accuracy of recognition. And this can also bring us a more realistic test environment. On the other hand, with the development of science and technology, the local speech recognition accuracy will increase, which will also bring an increase to the recognition accuracy of fraudulent calls for our application.

Conclusion

In this paper, we proposed a method to detect fraudulent calls by understanding the content of calls. First, this research collected textual data of telecommunication fraud on the Internet. Then, this research analyzed the textual data using machine learning algorithms and selected high-quality dataset from Sina Weibo data and Baidu data. On the next stage, this paper designed a method to extract features of words from textual data by using Natural Language Processing (NLP) techniques. Then we built rules for telecommunication fraud detection. Finally, we developed an application for telecommunication fraud detection on Android platform. The method of telecommunication fraud detection put forward by this paper break the limit of labeling caller number which is mainly used by most companies. The experimental results of this article show that the method proposed in this paper can be applied to practical applications.

Acknowledgements
We thank Zhen Wang for assistance with the technique of Android. And we are also grateful to Yue Zhao for her comments that greatly improved the research. We would also like to show our gratitude to Xuejing Yuan for sharing her wisdom with us in our research. And we thank "anonymous" reviewers for their insights. Besides, IIE authors are supported in part by National Key R&D Program of China (No.2016QY04W0805), NSFC U1536106, 61728209, National Top-notch Youth Talents Program of China, Youth Innovation Promotion Association CAS and Beijing Nova Program.

Funding
Our research was supported by National Key R&D Program of China (No. 2016QY04W0805), NSFC U1536106, 61728209, National Top-notch Youth Talents Program of China, Youth Innovation Promotion Association CAS and Beijing Nova Program. And the recipient is Professor Kai Chen.

Availability of data and materials
The manuscript is original work of all authors. We confirm that this manuscript has not been published elsewhere and is not under consideration by another journal.

Authors' contributions
All authors have contributed to this manuscript and approve of this submission. QZ participated in all the work and drafting the article. Prof. KC made a decisive contribution to the content of research and revising the article critically. TL has made many contributions to the technical route, designing research, and revising the article. YY participated in the performance detection and revising the article. XW made a decisive contribution to the idea in the early stages of his research and important intellectual contributions.

Ethics approval and consent to participate
We confirm that we have given due consideration to the protection of intellectual property associated with this work. We confirm that we have followed the regulations of our institutions concerning intellectual property. And all authors agree with its submission to Cybersecurity.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]SKLOIS, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China. [2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China. [3]Peking University, Beijing, China. [4]Indiana University Bloomington, Bloomington, USA.

Received: 13 April 2018 Accepted: 27 July 2018
Published online: 31 August 2018

References

2015 China Mobile Internet Users' Network Security Status Report, China Internet Network Information Center (CNNIC) (2016). http://www.cac.gov.cn/files/pdf/cnnic/2015phone.pdf

Abdallah A, Maarof MA, Zainal A (2016) Fraud detection system: A survey. J Netw Comput Appl 68:90–113. https://www.sciencedirect.com/science/article/pii/S1084804516300571

Azad MA, Morla R (2011) Multistage spit detection in transit voip. Software, Telecommunications and Computer Networks (SoftCOM), 2011 19th International Conference on. IEEE, pp 1–9

Azad MA, Morla R (2013) Caller-rep: Detecting unwanted calls with caller social strength. Comput Secur 39:219–236

Baidu (2017), Telecommunication fraud, https://www.baidu.com/

Blockeel H, Schietgat L, Struyf J, Džeroski S, Clare A (2006a) Decision trees for hierarchical multilabel classification: a case study in functional genomics. In: Proceedings of the 10th European conference on principles and practice of knowledge discovery in databases, pp 18–29

Chen KH, Chen HH (2001) Cross-language Chinese text retrieval in NTCIR workshop: towards cross-language multilingual text retrieval. ACM SIGIR Forum 35(2):12–19

Delamaire L, Abdou H, Pointon J (2009) Credit card fraud and detection techniques : a review. Banks Bank Syst 4(2)

Facts (2017) Chinese Investigation Report on Telecommunication Fraud Situations, vol 12 http://news.qq.com/cross/20170309/49rpD72V.html

Gao J, Li M, Huang C-N (2003) Improved source-channel models for Chinese word segmentation. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, Sapporo, pp 272–279. https://doi.org/10.3115/1075096.1075131

Jackson P, Moulinier I (2007) Natural language processing for online applications: Text retrieval, extraction and categorization, vol 5. John Benjamins Publishing

Jiang N, Jin Y, Skudlark A, Hsu W-L, Jacobson G, Prakasam S, Zhang Z-L (2012) Isolating and analyzing fraud activities in a large cellular network via voice call graph analysis. In: Proceedings of the 10th international conference on Mobile systems, applications, and services. ACM, pp 253–266

Jiang N, Jin Y, Skudlark A, Zhang Z-L (2013) Greystar: Fast and accurate detection of sms spam numbers in large cellular networks using gray phone space. USENIX Security Symposium, pp 1–16

Kolan P, Dantu R, Cangussu JW (2008) Nuisance level of a voice call. ACM Trans Multimed Comput, Commun Appl (TOMM) 5(1):6

Leontjeva A, Goldszmidt M, Xie Y, Yu F, Abadi M (2013) Early security classification of skype users via machine learning. Proceedings of the 2013 ACM Workshop on Artificial Intelligence and Security, ser. AISec, p 13

Li H, Xu X, Liu C (2018) A Machine Learning Approach To Prevent Malicious Calls Over Telephony Networks. 39th IEEE Symposium on Security and Privacy. IEEE, pp 561–577

Li Y, Yuan H (2017) What is the road to preventing and controlling telecommunications fraud, Daily inspection, 02.23. http://www.spp.gov.cn/llyj/201702/t20170223_181874.shtml

Lookman Sithic H, Balasubramanian T (2013) Survey of insurance fraud detection using data mining techniques. Int. J. Innov. Technol. Explor. Eng 3:62–65

Miramirkhani N, Starov O, Nikiforakis N (2017) Dial one for scam: analyzing and detecting technical support scams. NDSS

Open platform from iFLYTEK (2017). http://www.xfyun.cn

Oren N (2002) Reexamining tf.idf based information retrieval with Genetic Programming. In Proceedings of SAICSIT 2002, pp 1–10

Patankar P, Nam G, Kesidis G, Das CR (2008) Exploring anti-spam models in large scale voip systems. In: Distributed Computing Systems. ICDCS'08. The 28th international conference on. IEEE, pp 85–92.

Preacher KJ, Curran PJ, Bauer DJ (2006) Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. J Educ Behav Stat 31:437–448

Rebahi Y, Sisalem D (2005) Sip service providers and the spam problem. In: Proceedings of the 2nd VoIP security workshop

Rebahi Y, Sisalem D, MageDanz T (2006) Sip spam detection. Digital Telecommunications, 2006. ICDT'06. International conference on. IEEE, pp 68–68

Rebahi Y, Nassar M, Magedanz T, Festor O (2011) A survey on fraud and service misuse in voice over IP (VoIP) networks. Inf. Secur. Tech. Rep 16(1):12–19

Salton G, Buckley C (1988) Term-weighing approache sin automatic text retrieval. Inf Process Manag 24(5):513–523

Schwenk H, Gauvain J-L (2005) Training neural network language models on very large corpora, Proc. Joint Conference HLT/EMNLP

Sina Weibo (2017) Telecommunication fraud. http://s.weibo.com/weibo/.

Sorge C, Seedorf J (2009) A provider-level reputation system for assessing the quality of spit mitigation algorithms. In: Communications, 2009. ICC'09. IEEE international conference on. IEEE, pp 1–6

Srivastava K, Schulzrinne HG (2004) Preventing spam for sip-based instant messages and sessions

Subudhi S, Panigrahi S (2015) Quarter-sphere support vector machine for fraud detection in mobile telecommunication networks. Procedia Comp Sci 48: 353–359

Subudhi S, Panigrahi S (2017) Use of Possibilistic fuzzy C-means clustering for telecom fraud detection. In: Behera H, Mohapatra D (eds) Computational intelligence in data mining. Advances in intelligent systems and computing, vol 556. Springer, Singapore

Sun M, Li J, Guo Z, Yu Z, Zheng Y, Si X, Liu Z (2016) THUCTC: An Efficient Chinese Text Classifier.

Wang F, Mo Y, Huang B (2007) P2p-avs: P2p based cooperative voip spam filtering. In: Wireless Communications and Networking Conference, 2007. WCNC 2007. IEEE, pp 3547–3552

Wang F, Wang FR, Huang B, Yang LT (2013) Advs: a reputation-based model on filtering spit over p2p-voip networks. J Supercomputing:1–18

Wu Y-S, Bagchi S, Singh N, Wita R (2009) Spam detection in voiceover-ip calls through semi-supervised clustering. Dependable Systems & Networks, 2009. DSN'09. IEEE/IFIP international conference on. IEEE, pp 307–316

Zhang G, Fischer-Hubner S (2011) Detecting near-duplicate spits in voice ¨ mailboxes using hashes. ISC. Springer, pp 152–167

Zhang R, Gurtov A (2009) Collaborative reputation-based voice spam filtering. In: Database and Expert Systems Application, 2009. DEXA'09. 20th International Workshop on. IEEE, pp 33–37