Analysis of Telecommunication Fraud Cases Based on TF-IDF Algorithm

Hulin Yang and Ai’nan Zhu

People's Public Security University of China, Beijing, 100038, China

Abstract. [Purpose/Meaning] In the fast-developing information age, the "Internet +” model has deeply affected every bit of daily life. This has also created a lot of information and data for criminals who commit telecom fraud, leading to telecom fraud Become a new type of "non-contact" high-incidence crime. [Methods/Process] This study uses text analysis, using python crawler, Jieba word segmentation, TF-IDF algorithm and other tools to analyze and calculate the obtained text, using SPSS modeler to mine its data. [Result/Conclusion] Analysis of the content of the text obtained, found that as long as strangers on the phone there is a 40% probability; women become victims than twice as much as men; telecommunications fraud crime brush orders, fake customer service friends are still commonly used criminal means.

Keywords. Telecommunication fraud; Visualization; Information analysis; Data mining; Criminals

1. Introduction

Telecom fraud is a new type of high intelligence, non-contact crime with rising crime rate in recent years, which has exceeded the traditional crime cases.[1] With the deepening and application of the Internet plus mode, criminals have introduced advanced "crime" concepts and scientific and technological means. This is also an important reason for the rapid development of telecommunications fraud and rapid technological upgrading. At present, the research on Telecom fraud mainly focuses on the causes of Telecom fraud, investigation methods, evidence collection methods and how to prevent them. [2] This paper mainly uses Python crawler technology for text mining and analysis of the data, from the perspective of visualization and intelligence information analysis, the words often appear in the text and the fraud means often used by criminals are visually analyzed and clustered, which provides a new idea and perspective for the research of Telecom fraud.

2. Research design

2.1. Data sources

This article crawls through the "typical case of fraud cases” in WeChat official account sponsored by the Public Security Bureau of a certain province. Altogether, it obtains 225 cases, cleaning and dispelling the collected text data, using Jieba segmentation to extract and extract the high-frequency words, and making the telecommunication fraud word cloud chart, and then using the TF-IDF algorithm to select the high frequency words in the text content after cleaning. This paper analyzes 10 words with close research content, and uses hierarchical clustering analysis method to cluster the text content to obtain hierarchical clustering tree.

2.1.1. Jieba participle. Jieba word segmentation is a kind of word segmentation method suitable for Chinese word segmentation. It mainly supports three kinds of word segmentation modes: precise mode, full mode and search engine mode. It also supports traditional Chinese word segmentation,
user-defined dictionary and MIT authorization protocol. [3] This multi-mode word segmentation method makes the analysis of text content more objective and reasonable. In this paper, through the "Jieba word segmentation" to obtain the text content of high-frequency word statistics, a total of 2972 high-frequency words, through the elimination of some auxiliary words, modal particles and high-frequency words not related to this paper, finally get the top 20 high-frequency words, as shown in Table 1.

| High frequency words | criminal | transfer | loan | account | phone | internet | QQ/WeChat | bank | card | APP |
|----------------------|----------|----------|------|---------|-------|----------|-----------|------|------|-----|
| Rate                 | 678      | 332      | 318  | 257     | 208   | 154      | 149/104   | 136  | 114  |

2.1.2. Wordcloud visualization. Wordcloud is a visual display of words with high frequency in text documents. Wordcloud mainly sets the frequency of low frequency and high frequency to remove a large number of low-frequency and low-quality words and high-frequency words irrelevant to this study, so that readers can clearly understand the main content of the text documents by looking at the generated wordcloud map directly. And key information.[4]

2.2. TF-IDF calculation
TF (term frequency) is the number of times a word appears in a text document. Because the length of different documents is different, the frequency of these words is different, so we need to normalize them, so that these words can be compared in the same environment. In order to realize the normalization between different words, the most common way is the ratio between the frequency of words and the total number of words in the text document. [5] TF-IDF calculation is usually used to measure the importance of a word or a word in a text set to the text containing the word or word. [6] TF-IDF algorithm formula is as following Figure:

$$F(w) = TF(w) \cdot IDF(w) = TF(w) \cdot \log \frac{N}{n(w) + 1}$$

| Figure 1. TF-IDF algorithm formula. |

2.3. Hierarchical clustering analysis of subject words
Hierarchical clustering, also known as system clustering, refers to the process of clustering according to a certain level, clustering samples, so that samples with similar characteristics gathered together, so that samples with large differences separated. Euclidean distance is used to measure the sample distance, that is, the distance between two samples(x, y) is the square root (k variables) of the sum of the squares of the differences between the values of each variable of each sample.[7]

3. Modeling analysis
3.1. Multiple scatter plot
In this paper, based on the features extracted by TF-IDF algorithm, after cleaning and screening the data, the following indicators are selected for multiple scatter diagram analysis, and the analysis results are shown in Figure 2.
Figure 2. Multiple scatter plot.

In the whole scatter diagram, it can be found that the standardization results of keywords such as stranger, loan, customer service and transfer are the highest in Telecom fraud crimes, which also provides a new direction and focus for the prevention of Telecom fraud crimes. When receiving a call from a stranger or pretending to be a good QQ friend, if you pretend to be a good friend, customer service and other staff, you will be involved in the process of calling. The information such as payment, transfer, account and bank card can be identified as telecom fraud. The victim can hang up the phone or delete a friend or confirm whether it is a friend in time to avoid property loss.

3.2 Logistic regression model

Through the above TF-IDF algorithm to extract the feature words, after manual screening, the following indicators are finally formed. Through the construction of logistic regression model for the indicators, the construction results are shown in Table 2.

Table 2. Logistic regression analysis.

| Category       | B       | Std. Error | Wald | df | Sig. | Exp(B)   | 95% Confidence Interval for Exp(B) |
|----------------|---------|------------|------|----|------|----------|-----------------------------------|
|                |         |            |      |    |      |          | Lower Bound       | Upper Bound      |
| loan           | Intercept | 676.087    | 29893.755 | .001 | 1   | .982     | 7.608E-77         | 2.286E+74       |
|                | frequency | -2.026     | 88.391  | .001 | 1   | .982     | .132               |                      |
| phone          | Intercept | 1658.891   | 39128.570 | .002 | 1   | .966     | 4.050E-116        | 2.359E+110       |
|                | frequency | -5.779     | 132.616 | .002 | 1   | .965     | .003               |                      |
| Customer service | Intercept | 847.461    | 32662.584 | .001 | 1   | .979     | 1.245E-86         | 4.324E+83        |
|                | frequency | -2.612     | 99.589  | .001 | 1   | .979     | .073               |                      |
| refund         | Intercept | 9692.142   | 43419.532 | .050 | 1   | .823     | 1.610E-34         | .000              |
|                | frequency | -77.811    | 411.938 | .036 | 1   | .850     | 9.914E-123        | 3.108E+116       |
| internet       | Intercept | 1761.307   | 39797.546 | .002 | 1   | .965     | 1.610E-34         | .000              |
|                | frequency | -3.435     | 140.094 | .002 | 1   | .964     | 9.914E-123        | 3.108E+116       |
| WeChat         | Intercept | 6605.869   | 34576.030 | .037 | 1   | .848     | 2.796E-19         | 5.184E-158       |
|                | frequency | -42.721    | 162.984 | .069 | 1   | .793     | 2.796E-19         | 5.184E-158       |
|          | Intercept  | frequency | p-value | t-value | CR   | z-score |
|----------|------------|-----------|---------|---------|------|---------|
| Bank     | 2784.175   | 41190.305 | .005    | 1.946   | 1.148E+13 |
| card     | -13.218    | 174.812   | .006    | .940    | 1.818E-6   |
| account  | 1517.721   | 37994.753 | .002    | .968    | 1.925E+104 |
| transfer | 46.852     | 10856.820 | .000    | .997    | 1.148E+143 |
| APP      | 6323.762   | 36277.715 | .030    | .862    | 5.610E+130 |
| QQ       | 2494.876   | 41027.911 | .004    | .952    | 1.642E+137 |

In this study, it is found that in many indicators, taking strangers as the benchmark category, the fraud methods are mainly impersonating friends, acquaintances, swiping orders, consulting and other methods; the fraud tools are mainly wechat, QQ and other dating software and other apps; in the process of fraud, it is mainly through the designated special account, bank card, transfer and other ways. Through the establishment of logistic regression model, it is found that the crime of telecommunication fraud has gradually evolved into a new kind of "group crime", with mature and complete criminal methods, strong pertinence and advanced criminal means. It is a "new type of crime" in the whole type of crime, and has gradually formed a complete upstream crime, namely "divulging citizens' information privacy" and downstream crime "money laundering", aiming at the crime Telecom fraud, an "industrial chain crime", not only needs to crack down on current crimes, but also needs to crack down on upstream and downstream crimes.

4. Conclusion

Through the above analysis, as a special "new type of crime", the crime of telecommunication fraud needs the participation of the whole people to combat and prevent. Based on the above analysis, in the prevention of Telecom fraud crime, we need to be alert to strangers' phone calls, wechat, QQ friends, etc. If the information such as loan, transfer, account and bank card is involved in the call process, it is likely to be Telecom fraud. It needs the parties and public security organs to intervene in the call, and it also needs the joint efforts of banks and other financial institutions to discover and crack down on the downstream crimes of Telecom fraud.

References

[1] Bai Lei Yu. Internet plus background investigation of telecommunications fraud cases[D]. Chinese People's Public Security University, 2017.
[2] Hu Xiangyang, Liu Xiangwei, Peng Wei. Research on prevention and Control Countermeasures of Telecom fraud[J]. Journal of people's Public Security University of China (Social Science Edition), 2010,26 (05): 90-98.
[3] Zeng Xiaojin. Implementation of Chinese stuttering segmentation technology based on Python[J]. Information and computer (theoretical Edition), 2019,31 (18): 38-39 + 42.
[4] Xu Bolong. Design and optimization of word cloud based on Jieba and wordcloud[J]. Fujian computer, 2019,35 (06): 25-28.
[5] Ye Xuewei, Mao Xuemin, Xia Jinchun, Wang Bo. Improvement of TF-IDF algorithm for text classification[J]. Computer engineering and application, 2019,55 (02): 104-109 + 161.
[6] Zhou Yuan, Liu Huai LAN, Du Peng Peng, Liao Ling. Research on text classification model based on improved TF-IDF feature extraction[J]. Information science, 2017,35 (05): 111-118.
[7] Chen Xiaohong, Wang Yujian, Wan Luhe, Xie Ruifeng. Study on ecological agriculture regionalization in Northeast China based on hierarchical cluster analysis[J]. Economic geography,
2012,32 (01): 137-140.