Pisco Med Publishing

A Research on Text Analysis of Telecommunication Network Fraud Information

Jiazhen Sheng, Weiwei Wang, Kuiyi Liu*

Southwest Minzu University, Chengdu 610041, China.

Abstract: With the advancement of Internet technology, the telecommunication network text information has featured novelty and diversification. The ever-changing text information involving cyber fraud poses new challenges to maintaining social order and protecting users' legitimate rights and interests. This paper first interprets the necessity of an analysis on the telecommunication network text information involving cyber fraud. Then, it analyzes the characteristics of this kind of information from its definition, dissemination mode and audience. Finally, it illustrates the application of text analysis on telecommunication network text information involving cyber fraud and the problems to be further studied from the aspects of Chinese word segmentation, sensitive word frequency, dictionary construction, etc., and explains some precaution measures.

Keywords: Telecommunication Network; Text Analysis; Cyber Fraud

1. Introduction

Informatization promotes the rapid increase of Internet and telecommunication network users. Each coin has two sides, and network development is no exception. The arrival of the information age is accompanied by huge negative effects [1]. In recent years, text information involving cyber fraud in the telecommunication network has emerged continuously. The non-contact dissemination of illegal frauds through the network, illegal links, SMS, new social media and other carriers show a high incidence. Therefore, it is of great significance to analyze and study the text information involving telecommunication network fraud and improve citizens' precaution consciousness, thus further creating a good network and social environment.

2. Characteristics of Telecommunication Network Fraud Information

2.1 Definition

Telecommunication network text information involving cyber fraud is usually expressed and disseminated by publishers in a targeted way. With communication as the media and computer network information system as the operating platform, these publishers disseminate fraudulent information such as fictional facts, malicious rumors, distorted truths, etc. to unspecified society groups, luring telecommunication network users to believe it, or even virtually mislead victims to become the disseminators of frauds. Fraudulent information in the telecommunication network is mainly disseminated by text. Generally, there are traditional and new modes of disseminating this kind of information.

2.2 Dissemination Mode

2.2.1 Traditional Mode

SMS is a common traditional channel for disseminating telecommunication network fraud information. Nowadays, almost everyone owns a mobile phone. People's daily life has always been influenced by all kinds of SMS. We need to log in to third-party software, register and handle all kinds of business through mobile phones and verify identification by SMS. SMS service is like a double-edged sword, which not only brings convenience to our daily life, but also hides all kinds of information involving cyber fraud.

2.2.2 New Mode

In recent years, more and more shifts from SMS fraud to cyber fraud exploiting network tools have emerged. The *2019 Research Report on the Trend of Cyber Fraud* published on January 7, 2020 by 360 Hunting Net Platform once showed that the top three main channels for victims to get in touch with fraudsters or fraudulent information were QQ, WeChat and telephone. The reports of the frauds through the three channels accounted for 10.69%, 10.38% and 9.76% of the total respectively [2]. New modes of disseminating fraudulent information, including those applying the latest communication tools and social software, are constantly emerging.

2.3 Audience

The audience of telecommunication fraud can be both young people who are active in social networking and the elderly in various ways. The elderly or people with low literacy seldom use the Internet, so they are ill-informed, credulous, and lack judgment. Besides, they are even misled to become disseminators to spread the fraud information in their own circles. Consequently, it's difficult for them to be aware of cyber fraud. Usually, they will not find the fraud until they have spread the fraud information and suffered losses. Potentially, this is also an objective reason for the coexistence of new and traditional telecommunication network frauds [3].

3. Text Analysis on Telecommunication Network Text Information Involving Cyber Fraud

3.1 Chinese Word Segmentation

Word is the smallest, most isolated and most meaningful unit of a text. Word segmentation is the basis of phrase division, concept extraction, hotspot analysis and feature understanding for all kinds of texts, including telecommunication network texts involving online fraud. Simply put, Chinese word segmentation is to divide continuous Chinese character strings into individual words. Essentially, word segmentation is a process of re-segmenting successive Chinese character sequences according to the norms of Chinese words and then composing word sequences. Hence, word segmentation is also the key of text analysis involving telecommunication network fraud. Currently, there are three main word segmentation methods in the research field: ① word segmentation based on statistics ② word segmentation based on dictionary ③ word segmentation based on understanding [4].

Obtaining sensitive words involving cyber fraud is the first step for identifying the telecommunication network text information. The fundamental task of text analysis on telecommunication network text information is to quickly identify new topics, sensitive topics, emergencies, etc., from massive text data. The basic unit of text information is a single word, which is acquired by word segmentation algorithm.

3.2 Sensitive Word Frequency

Finding out the word frequency is a simple way to judge the importance of a word to the text. The judging index of word frequency statistics can be expressed by the ratio between the number of occurrences of a word to the sum of the occurrences of all words in the text. However, the main sources of telecommunication network text information are SMS, WeChat, QQ, MMS and network links. These data are all composed of short texts. Hence, in an identification with the word frequency, the words that appear most often are common words with no practical meaning, such as "yes" and "of". Therefore, before the word frequency statistics, we should first build a text data set based on a certain scale of information data, and use the corresponding algorithm to calculate the word frequency. TF-IDF is exactly an applicable method.

TF-IDF, a frequency-inverse document frequency algorithm, is a statistical method for evaluating the importance of a term to a document in a file set or a corpus.[5] TF (term frequency) refers to the number of times a specified term appears in a document. The normalization formula of TF is:

$$\mathrm{TF} = \frac{\text{Number of times term appears in a document}}{\text{Total number of terms in the document}} \quad (1)$$

As for IDF, if there are fewer documents containing a certain term, the higher the IDF will be. And a higher IDF indicates that the term enjoys a good ability to classify. The IDF value can be worked out according to the formula below, in which “+1” is to prevent the denominator from being 0.

$$\mathrm{IDF} = \log_e \frac{\text{Total number of documents}}{\text{Number of documents with a certain term}+1} \quad (2)$$

The high-frequency terms in a document, and the low file frequency of these terms can produce highly weighted TF-IDF values, so TF-IDF tends to filter out common terms and retain important terms.

$$\text{TF-IDF}=\mathrm{TF} \times \mathrm{IDF} \quad (3)$$

It is more reliable to measure the importance of a word to a category than to simply measure its importance to a document. However, TF-IDF only considers the word frequency, but does not consider the location information of words. Since words in different locations have different discrimination abilities, words in different locations should be treated separately. Hence, this algorithm also needs to be improved.

3.3 Dictionary Construction

The dictionary used in most of the current research is How Net dictionary published by CNKI. The dictionary divides words into six categories: positive emotion, negative emotion, positive evaluation, negative evaluation, degree and level and proposition. On this basis, we should adopt manual reading and case screening to further refine the exclusive dictionary of sensitive words in telecommunication networks involving cyber fraud. This is because the information related to telecommunication network fraud generally includes prize winning, remittance, information interception, loan, shop recommendation, training and education, fund and stocks, etc.[6] The sensitive words it contains are mainly non-characteristic and conventional words, such as those related to user reply, money, telephone contact, bank, invoice, real estate, shops, training, education and so on. Therefore, we should further revise the conventional dictionary, delete some words that can express emotional tendency only in a specific context, and add some words that are ambiguous only in the case of telecommunication network fraud.

4. Conclusion

In recent years, with the development of information technology and application, new types of illegal and criminal activities, especially frauds, in telecommunication networks, are increasing day by day. Additionally, criminals still use traditional channels to disseminate fraudulent information. To address this issue, we should coordinate governance, keep pace with the times and constantly update our precaution measures.

Furthermore, due to the inevitable loopholes in the supervision of operators, it is difficult to effectively and accurately define and distinguish fraudulent information. Meanwhile, both the complicated industrial chains of disseminating fraudulent

information and hidden traps such as language snares, MMS pictures, harmful links pose great challenges to detecting and giving early warning of such illegal frauds. To this end, we not only analyze the dissemination characteristics and audience of telecommunication network fraud texts, but also formulate targeted treatment methods from the perspective of text analysis. On this basis, the combination of machine learning, emotion analysis and image text recognition will be a promising orientation for further research [7].

References

[1] Lou Yongtao, TANG Xiang. *The prevention, control and reflection of telecom network fraud crime in the big-data age.* Journal of Chongqing University of Technology (Social Science) Vol. 2020, 34(03):121-128.

[2] *2019 Research Report on the Trend of Cyber Fraud*.

[3] Zhang Qi. *On the Approaches to Network Fraud Education for the Elderly from the Perspective of CIP Theory ———Based on "Mobile Payment Usage in China" from 2014 to 2020*. Contemporary Continuing Education, 2021, 39(03):74-80.

[4] Li Jia. *A Tentative Study on Chinese Segmentation Algorithm*. Software, 2013, 34(07):75-76+120.

[5] Yan Hanbing, Zhou Hao, Zhang Honggang. *Automatic Malware Classification via PRICoLBP*. Chinese Journal of Electronics, 2018, 27(04):852-859.

[6] Hao Wenjiang, Xu Liping, JIANG Jinlei, et al. *Research on Control Technology of Telecom Network Fraud Crime*. Netinfo Security, 2016(9): 213-217.

[7] Bing Wu. *Application of Adaboost Algorithm and Immune Algorithm in Telecommunication Fraud Detection*. International Conference on Network, Communication, Computer Engineering (NCCE 2018). 2018:5.