ORIGINAL RESEARCH

ADHYAYAN Volume 13, Issue 1, 2023 Print ISSN: 2249-1066 Online ISSN: 2455-8656

Identifying Fraud Detection Techniques Using Text Analytics Processing

Rajeev Tripathi[1], Smita Tripathi[2]

[1]Department of Computer Application and Sciences School of Management Sciences, Lucknow, Uttar Pradesh, India

[2]Department of Computer Application School of Management Sciences, Lucknow, Uttar Pradesh, India

Abstract

To determine the fraud detection model, to illustrate how the fraud detection model is created, and to start the data model with any classifier, data mining technology is used in the fraud detection process. As e-commerce continues to grow, the associated internet hoax is still a very appealing source of cash for scammers. Because of the severe financial damage that this counterfeit activity does to retailers, online fraud detection is essential. Concerned with scam detection is the need to quickly seize fraudulent actions in addition to containing them. This significance is essential to reducing financial losses. Cybercrimes are a widespread annoyance and have a negative impact on our society in many ways. In every nation, the police enforcement system heavily relies on the results of cybercrime investigations. The connected online hoax is still a highly alluring way for con artists to get money as e-commerce expands. Online fraud detection is crucial due to this counterfeit activity's significant financial harm to merchants. In addition to curbing fraudulent acts, scam identification raises the urgent need to seize them. To minimize financial losses, its relevance is crucial. Cybercrimes are a common irritant that harms our society in a variety of ways. The outcomes of cybercrime investigations substantially influence the police enforcement systems in every country. These methods are essentially utilized for fraud detection in a variety of industries, including health, insurance, online shopping, and more.

Keywords: Text Analysis, Taxonomy, Clustering data mining, Fraud detection.

Adhyayan: A Journal of Management Sciences (2023); DOI: 10.21567/adhyayan.v13i1.02

Introduction

The unauthorized conversion of another person's property for one's own use and advantage is a common definition of financial crime as a crime against property. Gaining access to and control over someone else's property for financial benefit is considered a profit-driven crime. Fraud is a common element in financial crimes. They interchangeably use the phrases financial crime, white-collar crime, and fraud.1 Check and credit card fraud, mortgage fraud, medical fraud, corporate fraud, bank account fraud, point-of-sale fraud, currency fraud, and health care fraud are all examples of financial crime. These crimes also involve insider trading, tax evasion, kickbacks,

embezzlement, identity theft, cybercrime, money laundering, and social engineering.2 Concentrating on financial crimes, including payment cards, money laundering, fake money, security papers, and social engineering fraud. According to the group, unchecked financial crimes will have an adverse effect on people, businesses, organizations, and even whole countries due to the financial losses they cause. In recent years, the importance of investigating and studying financial crimes has grown in both the corporateand governmental sectors, including academia and the government's enforcement agencies. There have been several studies on financial crime and white-collar crime undertaken in the private and governmental sectors. There are many different study areas,and the results are quite helpful in preventing scams. Financial crime arises in a wide variety of forms when new information about them and related academic disciplines become available. The classification and categorization of

Corresponding Author: Rajeev Tripathi, Department of Computer Application and Sciences School of Management Sciences, Lucknow, Uttar Pradesh, India, e-mail: rajeev@smslucknow.ac.in

How to cite this article: Tripathi, R., Tripathi, S. (2023). Identifying Fraud Detection Techniques Using Text Analytics Processing. Adhyayan: A Journal of Management Sciences, 13(1):5-8.

Source of support: Nil

Conflict of interest: None

financial crime are offered in information technology as semi-structured data and explained in plain language. Since the semi-structured material is taxonomically organized, it also cannot bebest kept and managed in a machine- or computer-readable format. As a result, researchers are interested in how to organize and present financial crime data and natural language (research writing and text) into formal language (machine-readable) so that computers can understand. The objective of this paper is to clarify the general research questions and terminology that are frequently used in the financial crime areas. The importance of this research lies in providing a broad overview of financial crime researchand in its ability to organize domain knowledge into a machine- or computer-readable format that will aid in the development of a financial crime-related information management system, such as fraud detection and prevention system, in the near future. The process of comprehending and representing specific domain information is known as knowledge representation in knowledge management. These statistics conceal important information and intriguing patterns. The potential for banks to use data mining in their decision-making processes in areas like marketing, credit risk management, money laundering, liquidity management, investment banking, and timely identification of fraudulent transactions is enormous. Failures in these areas might have unfavorable effects on the bank, including client loss to rival businesses, monetary loss, damage to the bank's brand, and significant fines from the authorities.

Background and Related Work

VISA and MasterCard have introduced several security measures at the transactional level that are either a result of the PCI-DSS (Payment Card Industry Data Security Standards) set of guidelines or the necessity of utilizing the new chip cards that employ EMV (Euro pay MasterCard VISA) technology[3]. It addresses a number of security flaws in previous magnetic strip cards. While EMV made great strides toward contactlessoperations, the majority of cards in Europe are contact-based, creating some of these cards.

Skimming vulnerabilities that take into account the type of card fraud that involves the theftof credit card information used in an otherwise legitimate transaction. With ISO 14443 becoming a payment standard, businesses can now take contactless payments from various card issuers, including Visa and MasterCard, which offers a great deal of easeto business owners

A comparison study of various biometric authenticators that could be used for online banking [8] showed that the fingerprint, iris, and face are the most suitable biometric authentication methods for online banking, particularly when two-factor authentication is required.4

In order to make the creation of a cardholder's spending profile easier, credit card transactions were trained using the Baum-Welch algorithm by modeling the sequence of operations using a Hidden Markov Model (HMM). Transactions were divided into three groups based on transaction amount: high, medium, and low[5].

To increase the security of newly occurring behaviors, a hybrid method for detecting credit card fraud was created. It combines the Naive Bayes algorithm with the Hidden Markov model.6

The goal of the principle component analysis, which was first proposed, was to represent each sample of transaction with a small number of values, allowing for faster identification of fraud in credit card transactions by reducing the number of attributes anddetermining which attributes contain the most important data[7].

The new web clustering: The LINGO method, which is based on Latent Semantic Indexing and Singular Value Decomposition and identifies good cluster labels with meaningful meaning before assigning the contents to each label, is believed to have a heavy focus on the high quality of group descriptions. When LINGO is applied into the Carrot2 framework, it generates properly specified and relevant clusters that have a considerable impact on the clustering quality.8

Although there have been numerous studies and efforts put into detecting credit card fraud, K-Mean was proposed as a clustering data mining algorithm to distinguish between legitimate and fraudulent transactions. Despite the apparent lack of real data, the proposed algorithm produced significant fraud detection results.9

By detecting fraud patterns specific to each client rather than looking for a general patternof fraudulent behavior, this effort attempted to improve the current Fraud Miner algorithm offered by the suggested new fraud detection approach[10].

Utilizing data pre-classification to distinguish between legitimate and fraudulent transactions for each client might speed up the fraud detection process since legitimate and fraudulent behavior varies over a longer length of time than legitimate behavior does. Comparing the Apriori algorithm to NB (Naive Bayes), SVM (Support Vector Machine), RF (Random Forest), and KNN (K-Nearest Neighbor) classifiers, it showed good performance in handling class imbalance and recorded the highest fraud detection rate. This method also recorded the highest fraud detection rate and enabled real-time detection.

This problem was addressed by introducing Lingo as a fundamentally different approach to finding and describing groups that sought to find meaningful cluster descriptions[8] before assigning snippers to them and introducing DCF(Description ComesFirst) method using Singular Value decomposition. It was observed that most clustering mining algorithms made the discovery process first before basing it on contentslabels inducted that occasionally resulted in some groups' descriptions being meaningless[9].

- Data pre-processing using text filtering, stemming, and stop-word removal.
- Feature Extraction, which sought to identify common objects and expressions.
- Cluster Label Induction, which identifies the optimal phrase for matching.
- Cluster Content Discovery, which allocates contents to the clusters that are produced.
- Apply cluster merging and calculate cluster scores for the final cluster creation.

Methodology

The section explains the approach for utilizing text analysis to create a fraud detection that represents financial criminology expertise. The research process has been broken down into the following stages:

Data Preparation

Before beginning the data preparation phase, a transactions simulator that is in charge of simulating transactions and preparing an appropriate imbalanced dataset is required due to the sensitivity and confidentiality of the needed cardholder data required for the test as well as the limitations of the banks to provide this data for the test.

Following the creation of the dataset, the following data pre-processing steps would need to be taken:

- Refine the data by removing the transactions related to customers who have just onetransaction in the dataset.
- Classify transactions as either legitimate or fraudulent.

Text Mining and Analysis Implementation with RapidMiner Text ProcessingExtension

The machine learning processes used in RapidMiner's data mining and text processing include Process Document From Files, Tokenize, Transform Cases, Filter Stop Words, Generating nGrams, and Filter Token (By Length).

Process Document Files

To use RapidMiner Text Processing, you must first create a Process that serves as the main operator in all analytic operations.

Tokenization

The tokenize operator will divide a document's text into a series of tokens. There are several ways to separate points, including using just non-letter characters, specifying characters, regular expressions, linguistic sentences, and linguistic tokens. In this study, we selected all non-letter characters to extract tokens made up of a single word[12].

Transform Cases

The next operator is called Transform Case, which changes all of the characters in a document to either lower case or upper case, depending on the context. In this instance, we change every character to lowercase.

Filter Stopwords

The majority of English stop words like "a," "about," "above," "also," "all," "nearly," "alone," and "n" must be eliminated from the text in order for the text processing to accurately retrieve frequent terms in the field. The Filter Stop Words Operator completed its task by eliminating all tokens that are stop words from the built-in stop words list of RapidMiner.

Generate n-Grams

The next step is to produce n-Grams that turn a document's tokens into term n-Grams. With this operator, n-Grams of tokens are created in a document. A sequence of n consecutive tokens of the same Length is referred to as a "n-Gram." The phrase "n- Grams," which is produced by this operation, refers to all sequences of n tokens in a row that are consecutive.11

Filter Token by Length

The research employs a Filter Token (By Length) to filter tokens based on character length. Minimum Characters and Maximum Characters serve as the operator's inputs. Only tokens or terms retrieved inside this range will be considered for developing the ontology.

Model Development

A term enumeration procedure is carried out to identify the terminology that is utilized consistently across all data sources. The RapidMiner Text Processing tools will carry out this operation using predetermined machine learning algorithms. By looking up the definitions of every phrase listed, the class may be found. The name of the class is determined to be the broadest phrase with meaning and potential to represent a certain category. The study then organizes and validates the instances of the categories into the classes or ideas previously identified by contacting the expert. A thesaurus and dictionary were also consulted to comprehend each term's definition and related ideas fully. This study used a top-down approach to class identification, starting with a generic class and moving toward more specialized classes.

Result and Discussion

The outcome of this investigation is described in this section. Based on this finding, the study was able to identify all of the common phrases and common knowledge that were utilized in all of the data sources as per required in the development process. The Risk Sector discusses how financial criminality could arise in several sectors, including the management, commercial, and economic ones. The action involved in the enforcement offinancial criminology is explained in the enforcement class. In the field of financial criminology, technology is frequently discussed in terms of IT services, systems, data value, and analysis. Financial criminology's Location and Time lessons outline the areas and times when financial crime typically occurs, whether on a state and national level, ona worldwide scale, or even in the virtual world. The sources of data and knowledge that the researcher learned while researching and studying financial criminology are described in the resources section. Last but not least, the Property class provides broad explanationsof several categories of property that relate to financial criminology.

Conclusion

The advantages of the domain representation include assisting researchers and investigators in comprehending the subject matter and central problems in financial criminology. Information technology professionals can create a knowledge base system of financial crime using ontology (Semi or Structured Knowledge). RapidMiner, a Big Data Analysis tool with extra text processing extensions, is used in the research's datamining and text processing processes[13]. To detect credit card fraud, many data mining approaches have been developed. Establishing fraud/legal patterns for each client makes it simpler to detect fraud by allowing customer profiles to identify typical and fraudulent actions on his account. The taxonomy is then constructed by the study using the terms and words in accordance with the techniques described in the methodology section. According to the findings, there are nine primary categories or ideas that are frequently studied in financial criminology, including people, places, times, things, offenses, offenses, risk sectors, laws, enforcement, technology, and resources.

References

Fridson, M. S. (2002). Financial Crime Investigation and Control (a review).

Gottschalk, P. (2010). Categories of financial crime. Journal of financial crime.

Greenemeier, L. (2006). Visa expands contactless card efforts. Information Week.

Parusheva, S. (2015). A comparative study on the application of biometric technologies for authentication in online banking. Egyptian Computer Science Journal, 39(4), 116-127.

Matheswaran, P., Siva, E., & Rajesh, R. (2015). Fraud Detection in Credit Card Using Data Mining Techniques. International Journal of Distributed and Parallel Systems, 2, 11-18.

Nancy, A. M., Kumar, G. S., Veena, S., Vinoth, N. A., & Bandyopadhyay, M. (2020, November). Fraud detection in credit card transaction using hybrid model. In AIP Conference Proceedings (Vol. 2277, No. 1). AIP Publishing.

Pawar, A. D., Kalavadekar, P. N., & Tambe, S. N. (2014). A survey on outlier detection techniques for credit card fraud detection. IOSR Journal of Computer Engineering, 16(2), 44-48.

Osiński, S. (2003). An algorithm for clustering of web search results. Master, Poznań University of Technology, Poland.

Vaishali, V. (2014). Fraud detection in credit card by clustering approach. International Journal of Computer Applications, 98(3), 29-32.

Seeja, K. R., & Zareapoor, M. (2014). Fraudminer: A novel credit card fraud detection model based on frequent itemset mining. The Scientific World Journal, 2014.

Sahri, Z., Shuhidan, S. M., & Sanusi, Z. M. (2018). An ontology-based representation of financial criminology domain using text analytics processing. International Journal of Computer Science and Network Security, 18(2), 56-62.

Tripathi, R., & Dwivedi, S. K. (2021). Identification of qr code perspective on enhancement of text mining approaches. International Journal of Research in Engineering and Science, 9(8), 47-52.

Dwivedi, S. K., Manna, M. S., & Tripathi, R. (2022). Interpretive Psychotherapy of Text Mining Approaches. In Cognitive Informatics and Soft Computing: Proceeding of CISC 2021 (pp. 297-308). Singapore: Springer Nature Singapore.