# ESCUELA POLITÉCNICA NACIONAL

## FACULTAD DE INGENIERÍA DE SISTEMAS



## FRAUD DETECTION BY ANALYZING HUMAN BEHAVIOR APPLY MACHINE LEARNING TECHNIQUES

**THESIS SUBMITTED AS PART OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF DOCTOR OF PHILOSOPHY IN INFORMATICS**

**MARCO POLO SÁNCHEZ AGUAYO**
marco.sanchez01@epn.edu.ec

**SUPERVISOR: PHD. LUIS FELIPE URQUIZA AGUIAR**
luis.urquiza@epn.edu.ec

**Quito, octubre 2023**


ESCUELA POLITÉCNICA NACIONAL



# TESIS

For the award of the degree of

## DOCTOR EN INFORMÁTICA

Resolution RPC-SO-43-No.501-2014

del Consejo de Educación Superior

Presented by

MARCO POLO SÁNCHEZ AGUAYO

Thesis supervised by

**LUIS FELIPE URQUIZA AGUIAR,**

**Professor of the Escuela Politécnica Nacional**

# FRAUD DETECTION BY ANALYZING HUMAN BEHAVIOR APPLY MACHINE LEARNING TECHNIQUES

Examen oral en

Oral examination by the following committee:

**Manuel Sánchez Rubio, Ph.D.**

Universidad de Alcalá, Universidad Internacional de la Rioja, External examiner.

**Ahmad Mezher, Ph.D.**

Universidad de New Brunswick, External examiner.

**Ana Fernanda Rodríguez Hoyos, Ph.D.**

Escuela Politécnica Nacional, Opponent member.

**Edison Fernando Loza Aguirre, Ph.D.**

Escuela Politécnica Nacional, Coordinator member.

**Julio César Caiza Ñacato, Ph.D.**

Escuela Politécnica Nacional, Internal examiner.

# STATEMENT

I hereby declare underpath that I am the author of this work, which has not previously been presented for obtaining any academic degree or professional qualification. I also declare that I have consulted the bibliographic references included in this document.

Through this declaration, I transfer my intellectual property rights corresponding to this thesis, to the Escuela Politécnica Nacional, as established by the Intellectual Property Law of Ecuador, its Regulations and the current institutional norms.

I declare that this work is based on the following articles of my authorship (as main author or co-author) related to the title of this thesis.

- **Sánchez-Aguayo** M, Urquiza-Aguiar L, Estrada-Jiménez J. Fraud Detection Using the Fraud Triangle Theory and Data Mining Techniques: A Literature Review. *Computers*. 2021; 10(10):121. https://doi.org/10.3390/computers10100121. **Journal SJR Q2**
- **Marco Sánchez**, Verónica Olmedo, Carlos Narvaez, Myriam Hernández and Luis Urquiza-Aguiar, "Generation of a Synthetic Dataset for the Study of Fraud through Deep Learning Techniques," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 11, no. 6, pp. 2534-2542, 2021, http://dx.doi.org/10.18517/ijaseit.11.6.14345. **Journal SJR Q3**
- **Sánchez-Aguayo M**, Urquiza-Aguiar L, Estrada-Jiménez J. "Predictive Fraud Analysis Applying the Fraud Triangle Theory through Data Mining Techniques". *Applied Sciences*. 2022; 12(7):3382. https://doi.org/10.3390/app12073382. **Journal JCR Q2**
- **Sánchez-Aguayo M**, Urquiza-Aguiar L. "Comparative Analysis of the Performance of Machine Learning Techniques Applied to Real and Synthetic Fraud-Oriented Datasets," *Springer International Publishing*, 2022, pp. 44-56, https://doi.org/10.1007/978-3/-031-18347-8_4.

- **Sánchez-Aguayo M**, Urquiza-Aguiar L. “Improving Fraud Detection with Semi-Supervised Topic Modeling and Keyword Integration”. *PeerJ Computer Science*. 2023; **Journal JCR Q2 *Under review***
- **M. Sánchez**, J. Torres, P. Zambrano and E. Flores, “FraudFind: Financial fraud detection by analyzing human behavior,” *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, 2018, pp. 281-286, http://doi.org/10.1109/CCWC.2018.8301739.

I also declare that I have acknowledged the collaboration of third parties, and the contribution made by other published or unpublished material.

**MARCO POLO SÁNCHEZ AGUAYO**

## CERTIFICATION

I certify that MARCO ROLO SANCHÉZ AGUAYO has carried out his research under my supervision. To the best of my knowledge, the contributions of this work are novel.

**PhD. Luis Felipe Urquiza Aguiar**

**ADVISOR**

## DEDICATION

*I dedicate this work to my beloved wife, Paulina, and our two wonderful children, Matías and Julian. I am deeply grateful to my family for your unwavering support and sacrifices to allow me to pursue my dream. I hope this accomplishment brings you great pride and happiness and that we may share many moments of joy and success. In particular, I want to express my heartfelt appreciation to my wife for her constant encouragement and unconditional love. Her unwavering belief in me has been a driving force throughout this journey. To my children, I want to thank you for your patience and understanding during this challenging time. I know that I have not always been fully present. Still, your love and understanding have been a source of motivation and inspiration for me.*

*Marco Sánchez*

## ACKNOWLEDGMENTS

*To my director and tutor, Luis Urquiza, for his support and confidence in me during the research process.*

*To José Estrada, research collaborator, who was ready to collaborate in everything in the most professional way.*

Thanks.

*Marco Sánchez*

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# RESUMEN

En la actualidad, estamos presenciando un aumento repentino en los casos de fraude en todo el mundo, lo que hace necesario adoptar estrategias proactivas para detectar indicios o sospechas antes de que se materialice el delito. Afortunadamente, el desarrollo continuo de tecnologías informáticas brinda oportunidades favorables para combatir y mitigar este problema. La minería de textos y el modelado de tópicos son herramientas eficaces que, junto con las teorías enfocadas en el análisis de este fenómeno, pueden ayudar a identificar temas relacionados con el fraude y descubrir actividades fraudulentas para tomar decisiones precisas. Dado que la información tangible sobre el fraude es limitada, se generaron varios conjuntos de datos sintéticos con frases basadas en la teoría del triángulo de fraude, que se utilizó para desarrollar el modelo propuesto. Para reconocer patrones en los documentos, se aplicaron técnicas de modelado de tópicos no supervisadas, seguidas de un enfoque semisupervisado que superó a los modelos no supervisados y proporcionó una interpretabilidad superior. Esto permitió establecer una relación entre los temas resultantes y los vértices del triángulo de fraude, y obtener altas probabilidades de que un documento pertenezca a un tema específico. Estas probabilidades se utilizaron para entrenar algoritmos de clasificación y predecir comportamientos sospechosos de fraude, con resultados prometedores. La evaluación de la aplicabilidad del modelo a todos los conjuntos de datos generados permitió determinar que el modelo era generalizable y, por lo tanto, útil para la detección de fraudes.

**PALABRAS CLAVE:** Fraud, Cybersecurity, Machine Learning, Topic Modeling, Human Behavior.

## ABSTRACT

Currently, we are witnessing a sudden increase in fraud cases around the world, which makes it necessary to adopt proactive strategies to detect indications or suspicions before the crime materializes. Fortunately, the continuous development of computer technologies provides favorable opportunities to combat and mitigate this problem. Text mining and topic modeling are practical tools that, together with theories focused on analyzing this phenomenon, can help identify fraud-related issues and discover fraudulent activities to make accurate decisions. Since tangible information on fraud is limited, several synthetic datasets with phrases based on the fraud triangle theory were generated, which was used to develop the proposed models. To recognize patterns in documents, unsupervised topic modeling techniques were applied, followed by a semi-supervised approach that outperformed unsupervised models and provided superior interpretability; This allowed for establishing a relationship between the resulting topics and the vertices of the fraud triangle and obtaining high probabilities that a document belongs to a specific topic. These probabilities were used to train classification algorithms and predict behavior suspected of fraud, with promising results. Evaluating the model's applicability to all the generated datasets allowed us to determine that the model was generalizable and, therefore, useful for fraud detection.


**KEY WORDS:** Fraud, Cybersecurity, Machine Learning, Topic Modeling, Human Behavior.

# PROLOGUE

Fraud has become a pervasive problem worldwide, affecting individuals, businesses, and governments. The ever-evolving techniques fraudsters use, and the increasing complexity of fraudulent activity pose significant challenges for those who detect and prevent it. While regulators and law enforcement agencies have put measures in place to address fraud, it remains a persistent threat to our security and financial stability.

In response to this challenge, scientists and researchers have turned to technologies such as machine learning and artificial intelligence to develop methods to detect them. In this thesis, we propose a promising approach that involves the application of topic modeling and theories such as the fraud triangle to detect behavior patterns related to this phenomenon. By analyzing datasets, keywords and topics associated with fraud can be found; these can then be used to build a model that can proactively find fraudulent activity. However, a significant barrier is the lack of datasets that can be used to verify and test such models.

To overcome this challenge, several synthetic datasets replicating real-world scenarios were generated and used to test the effectiveness of the proposed model. By doing so, we can gain valuable insight into the model's performance and refine it accordingly.

The proposed method is evaluated with these generated datasets, and its performance is validated. The analysis of several unsupervised and semi-supervised topic modeling approaches is carried out, compared, and the best one for this research is defined. Additionally, experiments are carried out with various classification methods, using the probabilities obtained by topic modeling in the different datasets. They are compared with deep learning algorithms to define the most appropriate technique to detect suspicious fraud behaviors.

This research and its results provide significant contributions to the fight against fraud. The approach can be further improved by incorporating additional data sources and developing more sophisticated algorithms. By taking advantage of the latest advances in science and technology, we can develop effective strategies to prevent and detect fraud, protecting our financial security and promoting stability and the common good. Ultimately, the goal is to create a comprehensive framework to detect and prevent fraudulent activity in real-time, contributing to a safer and more trustworthy society.

# 1 INTRODUCTION

Fraud in the business environment usually occurs due to deficiencies or omissions in the internal control systems. Companies usually take only some of the necessary measures to prevent this phenomenon. Unfortunately, they often find out very late after suffering embezzlement. People do not commit fraud if they are not given the opportunity to do so. One condition that favors this is the weakness or non-existence of internal control systems. The hierarchical level must also be considered a possible fraudster in a company since if it has sufficient access; it will have more possibilities to prosper. It is also essential to consider the possibility of collusion, one of the criminal forms that most violate an internal control system, which consists of carrying out irregularities through the agreement of two or more people, some of whom may be alien to the organization.

Fraud is a worldwide phenomenon that affects public and private organizations, including various illegal practices that involve intentional deception or misrepresentation. According to the Association of Certified Fraud Examiners (ACFE), fraud includes any intentional or deliberate act of depriving another of property or money by cunning, deception, or other unfair acts [1]. Moreover, fraud has increased considerably recently, affecting the interests of both financial institutions and their customers. A study conducted by Price Waterhouse Coopers found that 30 % of the companies they surveyed had already been victims of fraud. Moreover, 80 % of their fraud was committed within the companies'ranks, especially in administrative areas, such as accounting, operations, sales, and management level, without leaving aside the customer service dependencies [2]. Fraud-related activities, generally unknown within a company, determine a series of irregularities and illicit acts characterized by intentional deception committed by fraudsters. Most anomalies detected are due to the lack of internal control mechanisms. In such situations, scammers commit fraud by exploiting the weaknesses [3].

Since humans commit fraud, it is tightly coupled with human behavior. Thus, understanding

the motivations of perpetrators or their psychological and personality traits that drive them to cross ethical boundaries can provide a new perspective for fraud detection tools [4].

This chapter is organized as follows. First of all, Section 1.1 describes the problem statement. Then, in Section 1.2, the motivation that led to this research is presented. Section 1.3 establishes the objectives of the research. Section 1.4 describes the methodologies used. Finally, Section 1.5 shows the contributions of this thesis which are detailed, including how the problem is addressed in each of the articles that are part of the research and the publication status of each one.



## 1.1 PROBLEM STATEMENT

Fraud is considered a subset of internal threats, such as corruption, misappropriation of assets, and fraudulent declarations, among others [5]. In a more formal definition, fraud is "the use of one's occupation for personal enrichment through the misuse or deliberate misapplication of the resources or assets of the employing organization", according to the Association of Certified Fraud Examiners (ACFE) [6]. The ability to commit this activity is based on the need for more control mechanisms that institutions and companies have. In such circumstances, fraudsters commit fraud by taking advantage of these weaknesses.

There is a consensus that prevention should be a priority to minimize fraud through proper risk management. Avoiding fraud saves time and financial resources since detecting it after it occurs means the stolen assets are practically irrecoverable. To enhance fraud prevention, organizations should focus on the root of the problem by identifying the causes that lead people to commit fraud and to understand their behavior [7]. Many theories have attempted to answer this question. The most frequently cited in this context are Cressey's Fraud Triangle Theory (FTT) and Wolf and Hermanson's Diamond Fraud Theory (FDT) [8]. Both approaches analyze how perpetrators go so far as to commit fraud, which is discussed below.

Currently, there are different solutions [9] for detecting fraud, which is focused on using different tools that perform statistical and parametric analyses based on data mining techniques and analyses of behavior. However, none of them solve the problem of timely fraud detection [10].

When supported by data mining techniques, fraud analysis helps reduce the manual parts

of the detection/verification process. It makes the search for fraud more efficient. It is impossible to guarantee people's proper moral and ethical behavior, especially in the workplace. Due to this reality, a valid option for identifying possible evidence of fraud from available data is to use automatic learning algorithms. Many works cover fraud detection and use data mining techniques as the primary focus [11],[12],[13],[14]. Two issues of data-mining-based fraud-detection research are: the deficiency of the actual public data available in this domain for conducting experiments [15]—appropriate access to data for researching this area is extremely difficult due to privacy—and the lack of well-documented and published methods and techniques.



## 1.2 RESEARCH MOTIVATION

The number of individuals trying to cheat, steal, or defraud has increased considerably. From the point of view of the phenomenon of fraud, it is classified as internal and external. The external ones are exacerbated due to security breaches and the ease of access to information on methods and techniques that allow the confidentiality of data to be violated, accompanied by the rapid growth of the Internet, constantly new services, and the migration of personal data and reliable information. The cloud has allowed fraudsters to increase their skills and extend their reach to commit fraudulent acts. On the other hand, internal fraud encompasses a series of irregularities and illegal actions characterized by the intentional deception of fraudsters that leads to the misappropriation of money and other essential resources within an organization. Most known anomalies are due to weak internal control mechanisms. In such situations, fraudsters commit acts of fraud by exploiting these weaknesses. Fraud detection and prevention have become a challenge for companies trying to minimize incursions by fraudsters. Techniques constantly change, and companies must react and be one step ahead. Today, all large companies with sensitive information must have a fraud detection system. Standard fraud prevention techniques, such as password protection, are not enough. Different methods used for fraud detection have been proposed over the years and are usually within data mining and statistics. Still, artificial intelligence and machine learning are also valuable for the area.

Predicting fraud entails a series of challenges, which can be classified as operational and technological; the operational challenge relates to processes such as regulatory controls and how these evolve. Although it is essential to consider these challenges, they are outside

the scope of this research. On the other hand, the technological challenge considers the problems associated with the techniques and tools used to address this problem. The challenges related to this technical approach consider the information and its size since a large amount of data is generated daily. The models built must be adequate and fast to detect fraudulent transactions. It also considers unbalanced data, that is, the number of fraudulent transactions in a dataset is deficient, making it difficult to identify fraud. Data availability is critical in analyzing this problem since bank and user data are confidential and, therefore, difficult to obtain. Finally, data misclassification makes detecting and reporting fraudulent transactions difficult, and this causes misclassification of the models.

Based on the above motivations, this research evaluates the feasibility of proposing a method for effectively predicting fraud. Therefore, the research questions of this thesis can be stated as follows:

RQ1. What are the advances in fraud detection using topic modeling and machine learning techniques, and how have these techniques been applied to various fraud theories in recent literature?

RQ2. How can a model be developed through the topic modeling approach using fraud theories and machine learning that allows for effective fraud analysis?

RQ3. How can information related to fraud be obtained as test cases and training for pattern analysis and subsequent evaluation of learning algorithms?

## 1.3 OBJECTIVES

The main objective of this thesis is to investigate and propose a model that allows pattern recognition for the early detection of fraud. To achieve this goal, some specific objectives are:

- ❖ Review literature on fraud detection, analyzing works considering human behavior as a risk factor inherent to this problem.
- ❖ Analyze the different fraud theories and select the most appropriate to the field of study.
- ❖ Identify hidden patterns in a dataset that may be related to the selected fraud theory.

- ❖ Generate a synthetic dataset that fits the fraud theory.
- ❖ Validate the efficiency of synthetic versus real datasets.
- ❖ Define a model that predicts if a dataset's sentence belongs to one of the vertices of the fraud theory.

## 1.4 RESEARCH METHODOLOGY

The research presented in this thesis describes a three-phase approach aimed at improving fraud detection. In the first phase, "Identified Problem," an extensive literature review is conducted from two essential perspectives: the role of human behavior as a risk factor and the integration of Machine Learning Techniques with behavior-based fraud theories. The second phase, "Data Obtaining," explores synthetic data generation using deep learning techniques to overcome the challenge of obtaining sensitive fraudulent data. In the third and final phase, "Experimentation," a mechanism is proposed to detect fraud-related suspicious behavior. This mechanism combines topic modeling with a supervised classifier to identify potential fraud-related content. The study evaluates several text mining techniques on a fraud-related data set, selecting the most compatible one for fraud detection through topic modeling. The research then uses supervised machine learning models on synthetic data to determine whether a text can be identified as related to fraud.

A combination of two widely recognized methodological approaches has been adopted: Data Science Design and Data Mining Project Management (DSR and CRISP-DM), respectively. The DSR methodology focuses on designing and developing innovative solutions to practical problems in a specific context. In this case, using DSR allowed the development of a solution to the fraud detection problem. The DSR methodology is iterative and cyclical, so the solution is developed and evaluated throughout the research process. In this case, the feedback loop between design and evaluation allows the solution to be adjusted in response to the needs and demands that arise from experimentation [16].

On the other hand, the CRISP-DM methodology plays a crucial role in addressing data mining problems by providing a systematic approach to exploring and discovering patterns in data. By following the step-by-step process outlined in CRISP-DM, this research was able to identify meaningful patterns and gain deeper insights into the fraud detection problem. Moreover, the structured and formal framework offered by CRISP-DM ensures a clear un-

derstanding of objectives and guides the necessary steps to achieve them [17].

In the context of this study, the experimentation phase, which corresponds to phase 3 of DSR (Design and Development), involved implementing the CRISP-DM methodology and its six phases: 1. Business understanding, 2. Data selection and collection, 3. Data preparation, 4. Modeling, 5. Evaluation, and 6. Results. Adhering to this methodology produced an artifact as the output, providing a solid foundation for the subsequent evaluation stages [18].

In this sense, adapting the methodologies mentioned above, the research done in this thesis proposes 3 phases, as shown in Figure 1.1. The first phase (Problem identified), corresponding to "Business Understanding" of CRISP-DM, aims to collect the literature on fraud detection from two perspectives. On the one hand, studies that consider human behavior as a risk factor inherent to this problem are analyzed, mainly through theories related to fraud. Beyond exploring these theories, different works that use machine learning techniques for fraud detection are analyzed. In addition, papers that integrate ML techniques with behavior-based fraud theories, such as FTT and FDT, will be sought.

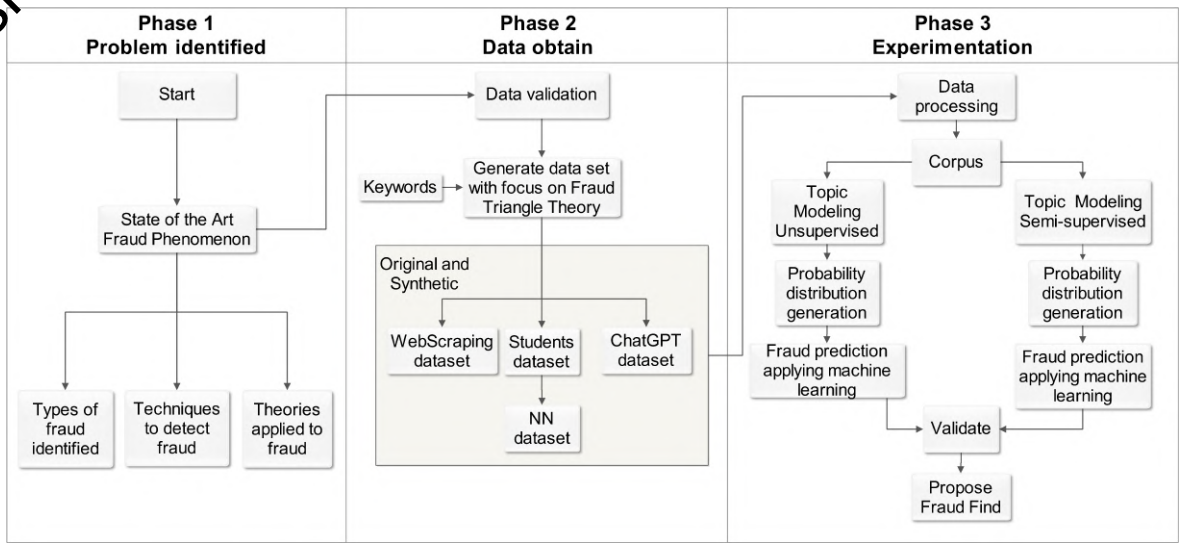


**Figure 1.1:** Methodology.

In the second phase (Data obtain), corresponding to "Data selection and collection" of CRISP-DM, the generation of synthetic data is considered a valid option for obtaining fraudulent data due to the difficulty of obtaining this information due to its sensitive nature. Therefore, deep learning techniques are discussed to show the application of commonly used algorithms to generate specific synthetic datasets practically and efficiently.

In the third phase (Experimentation), corresponding to "Data preparation and Modeling" of CRISP-DM, it is proposed to implement a mechanism to detect possible suspicious beha-

viors related to fraud by analyzing human behavior through PTT and leveraging machine learning (ML) and deep learning (DL) techniques, respectively. This detector will combine predefined topic modeling and a supervised classifier to alert potential fraud-related texts. To build our new detector, we will need to measure the performance of several commonly used text mining techniques that will be tested on the fraud-related dataset. Once the appropriate topical analysis technique is selected, we will use the probabilities of the documents on the assigned topic to determine if a text can be identified as fraud-related using supervised machine learning models using the synthetically generated dataset. The objective will be to show which technique is more compatible, working with topic modeling to detect behaviors suspected of fraud. This will allow us to identify the impact of using a particular dataset to analyze the effectiveness of the proposed model in detecting fraud. With these results and obtaining a performance baseline based on the different datasets, it is possible to examine the effectiveness of our proposed strategy on a classification problem involving fraud detection. The efficiency of other models of suggested topics (supervised) will be evaluated against the classic ones (unsupervised).

In this phase, the developed model's performance and effectiveness in fraud detection will also be assessed (Evaluation of CRISP-DM). This evaluation will involve analyzing its reliability and predictive power through performance metrics. The results obtained will provide valuable information about the performance of the model and its applicability in real-world scenarios.



## 1.5 CONTRIBUTIONS

In this section, we summarize the contributions of this thesis, specifying how the problem is addressed in each article and the publication status.

### 1.5.1 Chapter 2: Identify evidence from studies related to fraud.

This work aims to review current literature related to fraud detection that uses fraud theories, machine learning, and deep learning techniques. This systematic literature review must follow the methodological process illustrated in Figure 1.2, providing evidence that fraud is an area of active investigation. Several works related to fraud detection using machine learning techniques were identified without the evidence that they incorporated the fraud triangle as

a method for more efficient analysis. Beyond exploring these theories, on the other hand, our review analyzes different works where machine learning techniques have been used for fraud detection. Moreover, we look for works that integrate ML techniques with behavior-based fraud theories, such as the FTT and FDT.



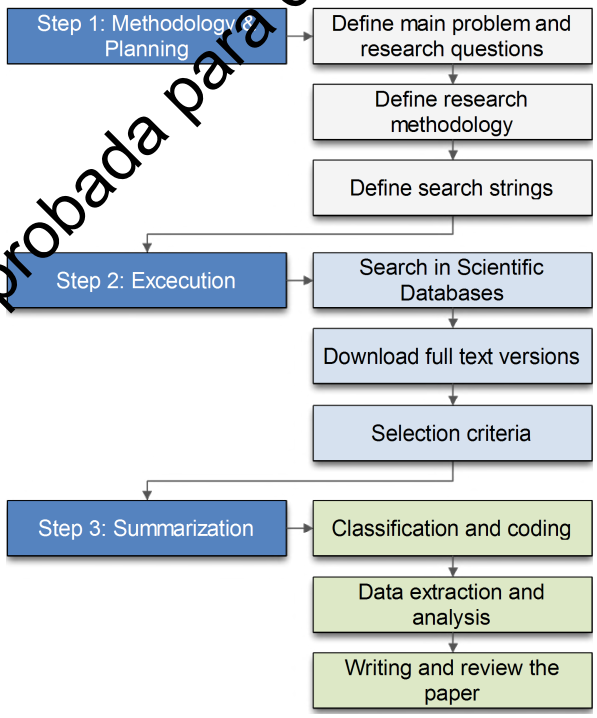


**Figure 1.2:** Methodology applied in the systematic literature review (SLR).

A total of 32 publications met all of the inclusion criteria. The selection of studies from the initial search identification phase and the final number of included studies are presented in Figure 1.3. As initially proposed and to ensure that the resulting reviews contained relevant information, we read the full text of the 32 studies to verify if they fit our adopted selection criteria. As a result, all of these publications represented our final set of primary studies.

Thus, this research aimed to identify publications related to fraud detection using ML techniques based on Fraud Theories. The proposed reference frameworks focus on developing tools that allow auditors to perform fraud analyses more efficiently by shortening their detection time through support from data mining techniques. In addition, the results of the quality evaluation carried out for the primary studies showed that the evaluation of their proposals was satisfactory in terms of the criteria of “relevance,” “limitations,” and “methodology.” When we assumed an approach to fraud detection through data mining techniques and using fraud theories associated with human behavior, this SLR reveals very little evidence from studies supporting this approach since only one preliminary study was found, corres-

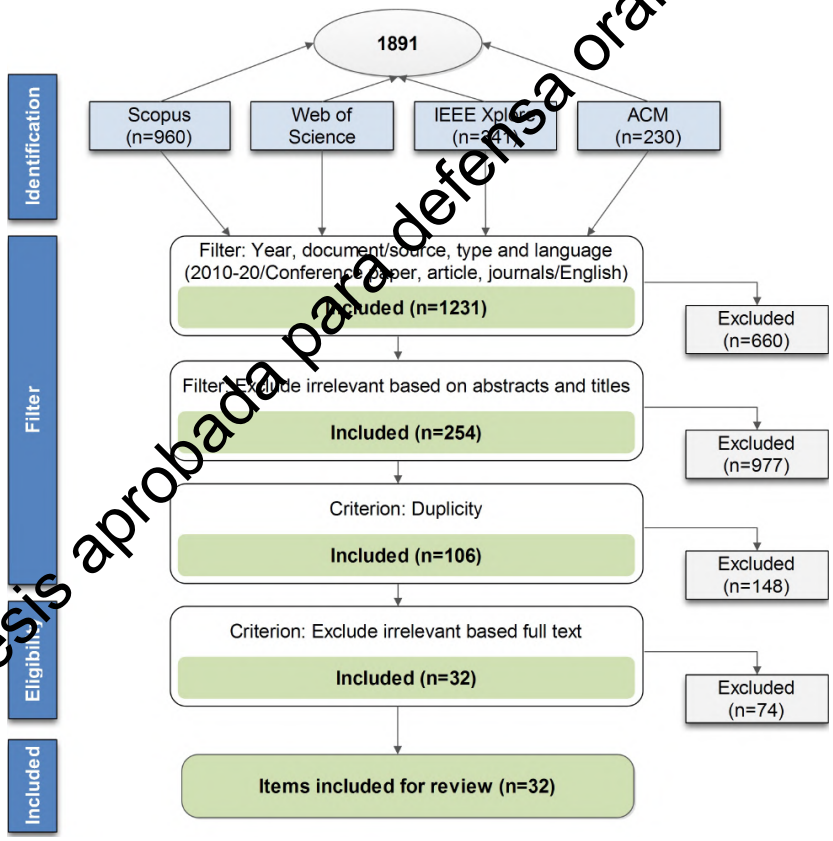


**Figure 1.3:** Steps followed to narrow the search results.



ponding to 3.13 % of the studies. When we allowed partial coverage, that is, fraud detection by applying only data mining techniques, 24 primary studies (corresponding to 75 %) could be classified. On the other hand, when we analyzed the approach to the analysis and detection of fraud in which only theories related to fraud associated with human behavior were considered, seven primary studies (corresponding to 21.88 %) supported this approach.

In this sense, only one study with evidence of the use of data mining techniques, the application of fraud theories, and a corresponding analysis of human behavior to detect fraud were identified, which means there is a gap, and this is an appropriate field to investigate.

This contribution was published in the **Journal:** *Computers; this article belongs to the Special Issue Artificial Intelligence for Digital Humanities-MDPI, 2021. SJR Q2.*

As a result of this SLR, the different contributions made about fraud in the scientific field were observed, identifying the required research gap. Therefore, in the subsequent investigation, it is proposed to build a model that allows the early detection of fraud and, above all, to obtain or generate a dataset associated with a theory of fraud that allows us to carry out the necessary experimentations.

### 1.5.2 Chapter 3: Model proposal to detect fraud from the context of human behavior.

This work proposes a mechanism to detect potential fraud by analyzing human behavior within a dataset. This approach combines a predefined topic model and a supervised classifier to generate an alert from the possible fraud-related text. Potential fraud would be detected based on a model built from such a classifier. As a result of this work, a synthetic dataset related to fraud is generated. In addition, our work suggests that this approach is feasible in practice since an average AUC performance more significant than 0.8 is obtained. Namely, the fraud triangle theory combined with topic modeling and linear classifiers could provide a promising framework for predictive fraud analysis.

The main contribution of this work is to propose a novel detector of suspicious behaviors related to the occurrence of fraud by analyzing human behavior using FTT leveraged on machine learning (ML) and deep learning (DL). Our detector combines a predefined topic model and a supervised classifier to alert a potential fraud-related text. More precisely, the suspicious phrases contain words that belong to a vertex of the fraud triangle (pressure, opportunity, and rationalization). On the other hand, non-fraudulent phrases have a general context that includes words unrelated to this problem. To build our novel detector, we have to do the following:

- Evaluate the performance of text mining techniques, such as Latent Dirichlet Allocation (LDA), non-negative matrix factorization (NMF), and latent semantic analysis (LSA) in the fraud-related dataset. The goal is to select the technique that provides an integral representation of the analyzed documents through clusters, i.e., topic, as separated;

- Once we select the appropriate topic analysis technique, we use the documents' probabilities on the assigned topic to determine if a text can be identified as fraud-related using supervised machine learning models. For this purpose, we conduct experiments on seven classification methods, including logistic regression (LR), random forest (RF), gradient boosting (GB), gaussian naive bayes (GNB), decision tree (DT), k-nearest neighbor (kN), and support vector machines (SVM), using the synthetically generated dataset.

- Furthermore, we perform the same experiment using deep learning techniques, such

as convolutional neural network (CNN), dense neural network (DNN), and long short-term memory (LSTM). To determine the performance's differences using receiver operating characteristic (ROC) curves based on the area under the curve (AUC) with the traditional ML classification methods. The goal is to show which technique is more compatible with topic modeling to detect suspicious fraud behavior.

The results can be seen in Figure 1.4. We compare the performance of linear classifiers and neural networks when applied to this scenario. The most efficient classification methods were RF and GB, averaging an AUC of 81 %.

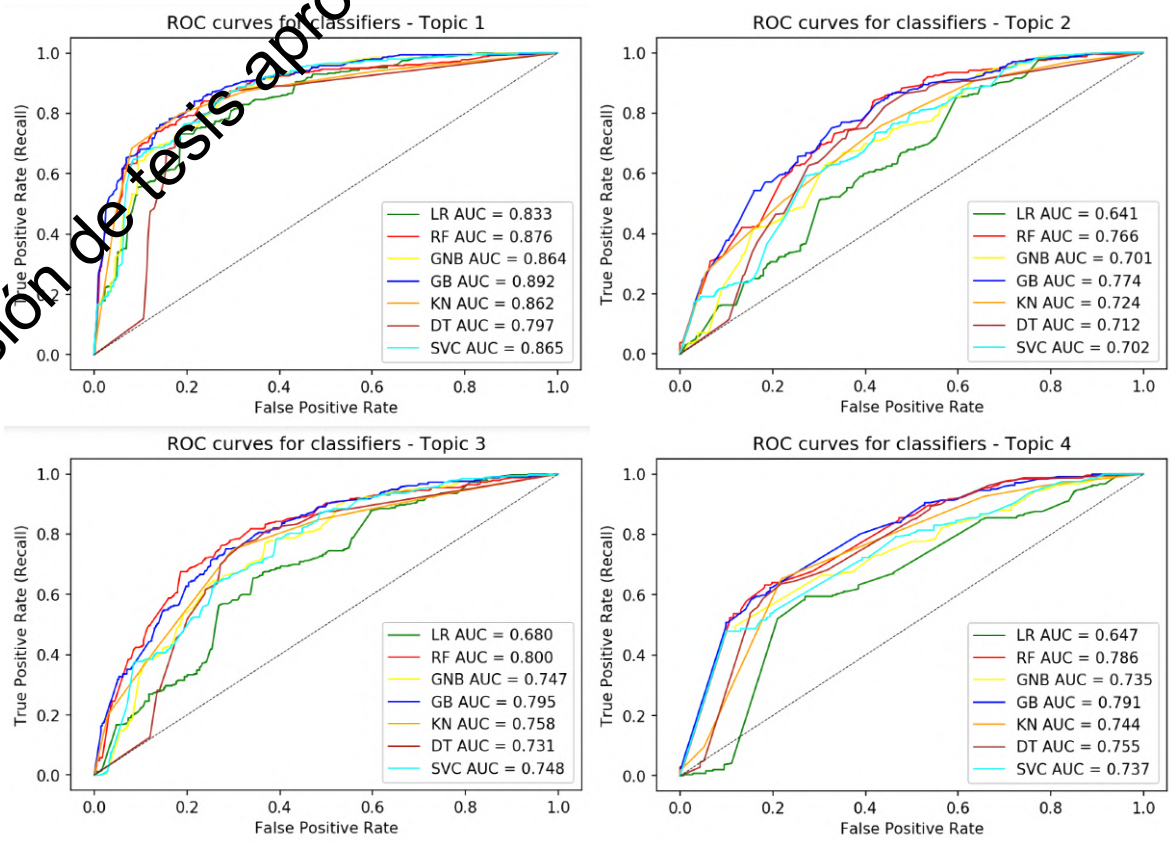


**Figure 1.4:** ROC curves of different classifiers for the datasets related to the dominant topics. SVC is the function in Scikit-learn, to implement SVM. (**a**) Topic 1. (**b**) Topic 2. (**c**) Topic 3. (**d**) Topic 4.

A graphical analysis of the inter-topic distance revealed that allocating documents to four topics resulted in a more coherent dataset interpretation. After assessing linear machine learning and deep learning algorithms, we found that some of the former were the best performers and obtained exciting results from AUC.

As noted, this work's novelty lies in combining a machine-learning mechanism with a sociological model to detect fraud-related behavior. As far as we know, such a model, the fraud triangle theory, is not used as a reference frame in any other work. Thus, our approach

might pave the way for addressing this problem from different perspectives, especially for incorporating other multidisciplinary approaches.

This contribution was published in the **Journal:** *Applied Sciences, 2022. JCR Q2.*

### 1.5.3 Chapter 4: Analysis of techniques and methodologies for the generation of synthetic datasets.

For an adequate analysis of fraud, it is necessary to have data that evidences this behavior. Even so, given that these data are scarce and difficult to find, generating synthetic data for their study is a viable option. We designed two algorithms to generate text to create a synthetic dataset for fraud analysis. The results obtained from this evaluation indicate that the data generation architecture proposed using the LSTM algorithm provides better performance in sentence readability (efficiency greater than 70 %) than RNN (less than 40 %). With LSTM, it was possible to synthesize a comprehensive dataset related to the fraud triangle's vertices.

The datasets obtained will be subjected to performance tests. Specifically, they will be compared using the Readability tool, which evaluates the text's coherence and provides score values between 0 and 100. The opposite occurs with satisfactory results of the LSTM algorithm (gray color). It surpasses the original data (blue color) with 77.71 % for Pressure. In comparison, Opportunity and Rationalization are above 70 %, as seen in Figure 1.5.

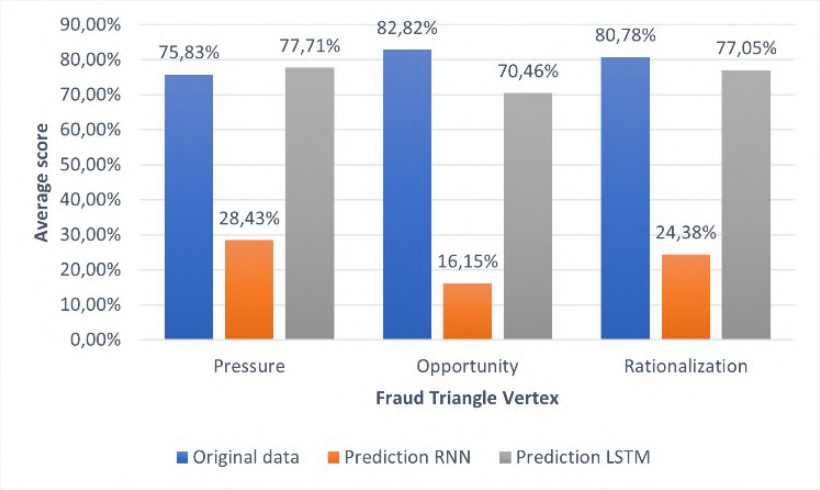


**Figure 1.5:** Comparison of averages obtained by the algorithms (RNN-LSTM) and the original data at the vertices of the fraud triangle.

This paper has presented a methodology for generating uniformly distributed synthetic data based on the fraud triangle theory. We compared the consistency of original and synthetically

generated data distributions based on their readability and grammar. Our results show that the original data's consistency score is higher than 70 %, which serves as the baseline. The synthetic dataset generated with the RNN algorithm is deficient and has a consistency below 40 %. On the contrary, the LSTM algorithm maintains a consistency level higher than 70 % and is similar to the original data's score.

This contribution was published in the **Journal:** *International Journal on Advanced Science, Engineering and Information Technology, 2021. SJR Q3.*

Three datasets have been generated, two synthetically and one real, using different methodologies. To identify if the use of these datasets has any impact on the model's performance, the performance of each of these should be analyzed, and the results compared.



### 1.5.4 Chapter 5: Study the applicability of various datasets generated synthetically in the proposed model.

For many organizations, sharing information is often a risk in terms of security and privacy, especially if the data is sensitive. In response to this problem, synthetic data emerges as a valid alternative, generated by different methods and techniques from an original or real dataset, allowing the sharing of information very close to reality. In this work, an experiment is carried out that validates the efficiency of synthetic versus real datasets by applying a model that predicts possible fraud cases in a dataset based on machine learning algorithms LDA and Random Forest or Gradient Boosting. We compared the prediction performance of our model over the real and synthetic datasets using metric ROC-AUC curves.

This work analyzes the validity of synthetic data generated through neural networks and tools [19, 20, 21] available on the internet, which synthesizes data based directly on real data of interest. The real data was obtained through simulation with students from the Escuela Politécnica Nacional (EPN). Validation of synthetic data for research requires comparing results derived from synthetic data with those based on original data.

Applying the model to several study datasets and examining the performance of the classifiers using ROC-AUC curves shows that in topics 0, 2, and 3, the RF and GB algorithms perform similarly with minimal differences. There are slightly more noticeable differences in theme one, but they do not significantly affect overall performance. For the real dataset "Students," RF and GB achieved an average AUC of 0.81. However, RF and GB had similar

behavior for the “WebScraping” synthetic dataset but achieved average AUC values of 0.81 and 0.83, respectively. On the synthetic dataset “Neural-Networks,” generated using deep learning, RF and GB reached average AUC values of 0.79 and 0.82.

The results suggest a similar behavior in the datasets analyzed based on the performance averages of the classifiers used, as seen in Figure 1.6.



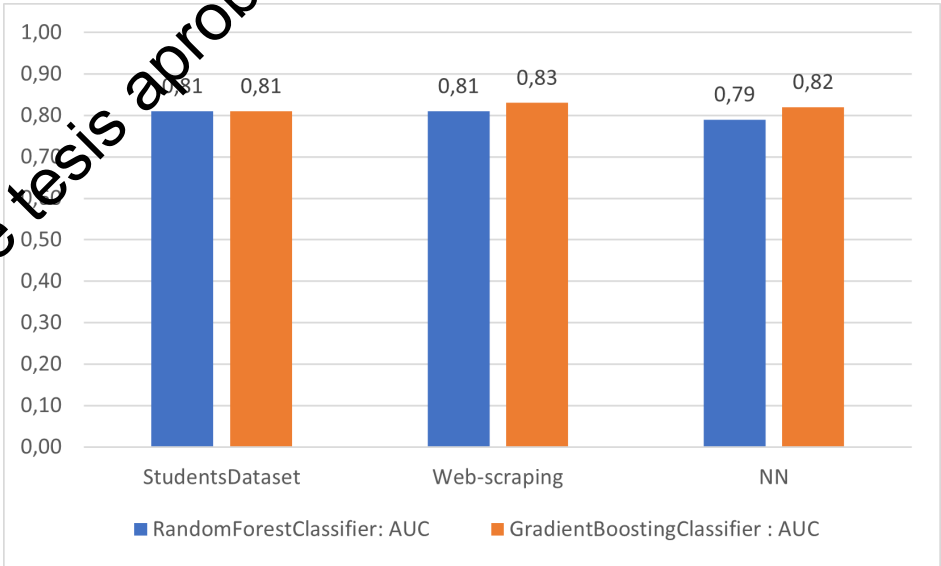


**Figure 1.6:** The best metrics obtained by the algorithms (Random Forest and Gradient Boosting) applied to the study datasets (Students, WebScraping, and Neural-Networks). The averages indicate a similar behavior in the analyzed datasets.

This work shows that the performance obtained by a detector of fraud-suspicious behavior based on machine learning algorithms used on the real dataset is similar to that obtained from synthetic datasets. These findings suggest that the results of models built using synthetic datasets may reflect behaviors obtained as if real data had been used. Synthetic datasets preserve the privacy and confidentiality of the information, allowing the development of predictive models to discover patterns without revealing confidential data, minimizing the risk of access to real data. In this case, according to the results obtained from the performance comparison, synthetic data is recommended to predict phrases suspected of fraud.

This contribution passed the first round of review in the **Conference:** *TICEC, 2022.*

### 1.5.5 Chapter 6: Topical Modeling Analysis with Semi-Supervised Approach and model validation.

Natural language processing techniques, such as topic modeling, have been explored to extract information and categorize large sets of documents. However, unsupervised topic modeling may not always produce the best results for specific tasks, such as fraud detection. Therefore, in the present work, we propose to use semi-supervised topic modeling, which allows the incorporation of specific knowledge of the study domain through the use of keywords to learn latent topics related to fraud. By leveraging relevant keywords, our proposed approach aims to identify patterns related to the vertices of the fraud triangle theory, providing more consistent and interpretable results for fraud detection. Overall, the study emphasizes the importance of deepening the analysis of fraud behaviors and proposing strategies to identify them proactively.

The main contributions are the following: first, we use CorEx as a topic model and perform an efficient alteration of its code to identify the probabilities that the corpus documents belong to a topic and to be able to visualize the distribution of topics through the pyLDAvis library. Second, we show how the fraud triangle theory can be integrated into CorEx through "keywords" related to the vertices of this theory. We show that CorEx produces more relevant topics than its unsupervised and semi-supervised variants of LDA.

Once the most efficient semi-supervised topic modeling has been identified, the probabilities that a document belongs to a specific topic are obtained, with which classification methods such as Gradient Boosting (GB) and Random Forest (RF) were trained to try to predict related cases with fraud. Finally, the proposed model is validated with the different datasets used in this research to try to establish the generality of the model.

Several synthetic datasets were used and generated to validate their performance to ensure the model's accuracy. The datasets were generated using various techniques to simulate different scenarios and environments [22]. The model was tested in multiple conditions to ensure it worked reliably in all situations, confirming that it could accurately predict outcomes in various contexts. The results of these tests were then used to validate the model's performance and provide evidence of its accuracy.

Implementing a predictive model aimed at detecting hidden patterns related to suspected fraud is the objective of the present study in which topic modeling techniques, including

LSA, NMF, and LDA, are used to identify which is the most effective, being LDA the chosen model. The number of topics is determined using a metric called coherence value (parameter k) to find the most appropriate number of topics, considering the nature of the data and the level of similarity between them. LDA helps identify representative words within topics and serves as a basis for semi-supervised learning algorithms to converge around these terms. The objective of topic modeling is to evaluate the probability that a document in a study corpus belongs to a specific topic associated with the aspects of the fraud triangle, as seen in the first phase of Figure 1.7. This process identifies possible fraud-related behaviors. The obtained probabilities are then used to train various classification methods to predict suspicious fraud-related activities. Evaluating these classifiers is essential to select the one most compatible with the analysis of the topic, ensuring the effectiveness and precision of fraud detection, as seen in the second phase of Figure 1.7.

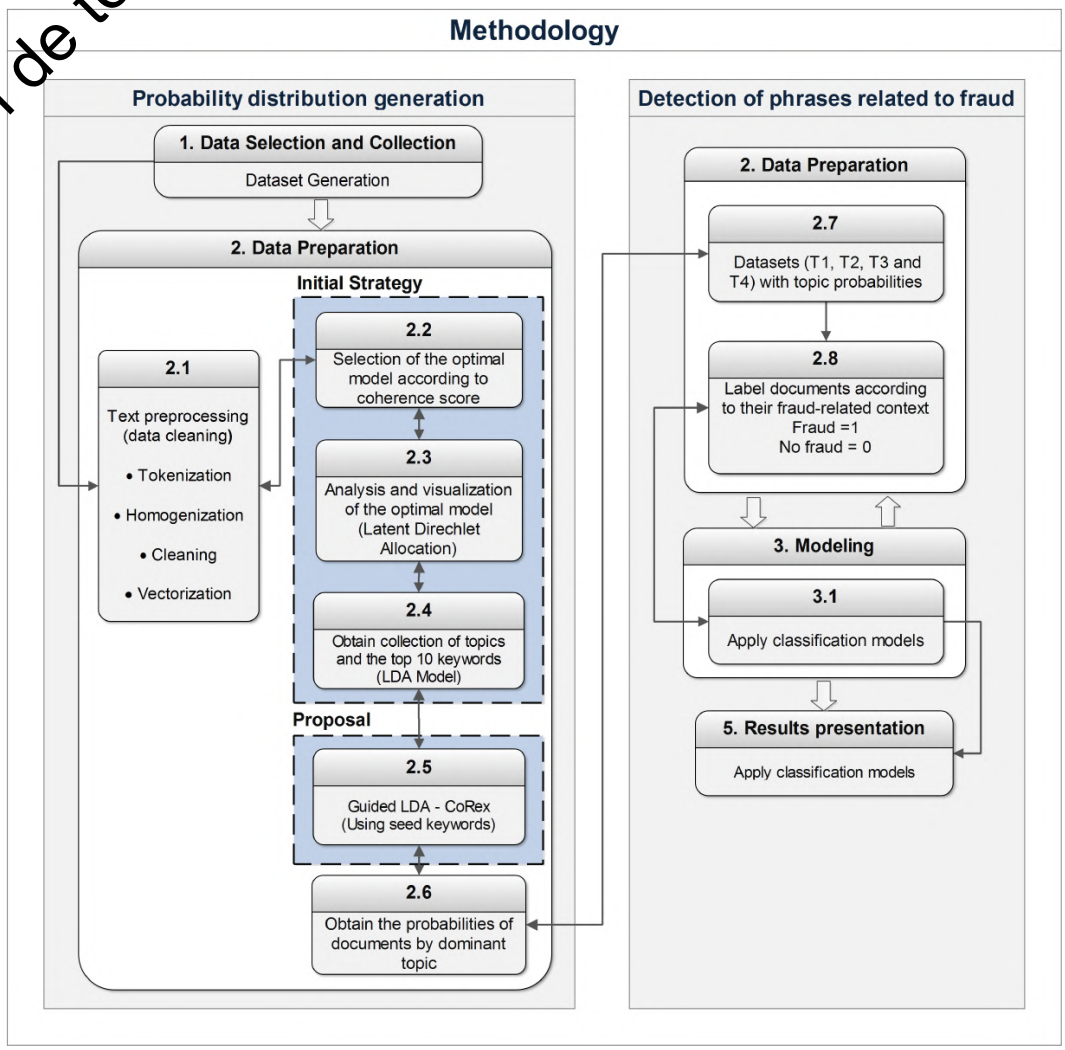


**Figure 1.7:** Methodology used to determine the existence of fraud.

This investigation applied topic modeling and machine learning techniques to analyze fraud-

related behaviors using a corpus of study and relying on the fraud triangle theory. The semi-supervised approach was applied in topic modeling; Corex and GuidedLDA algorithms were used, with Corex performing better in creating more coherent and interpretable topics aligned with the vertices of the fraud triangle. The probabilities of document-topic associations were extracted from the models and used as input for Gradient Boosting and Random Forest classification methods to predict fraud-related behaviors. Several datasets were used to test the model's performance. The results revealed good performance averages, indicating that the model can be generalized. It also showed that the model had a low bias and an acceptable variance in the four topics, indicating that it performed well on both the training and test sets.

This contribution was accepted and is in the final stages of review for publication in **Journal:** *PeerJ Computer Science, 2023. JCR Q2.*



### 1.5.6 Appendix: FraudFind: Financial Fraud Detection by Analyzing Human Behavior.

As a complement to the work carried out, a tentative approach is attached, which involves the development of a traffic capture and analysis tool to establish possible relationships with the concept of the "fraud triangle." It is important to note that this is the first approximation. However, it constitutes a starting point that still has room for considerable improvement by integrating the model that has been developed in the course of this research.

The present work proposes FraudFind, a conceptual framework to detect fraud supported by the fraud triangle factors. Compared to the classic audit analysis, it significantly contributes to the early detection of fraud within an organization. Considering human behavior factors, detecting unusual transactions that would not have been considered using traditional audit methods is possible.

The proposed framework is designed for continuous auditing in an organization, focusing on fraud detection. It incorporates the fraud triangle theory and considers the human factor as crucial. It aims to analyze large volumes of data from various sources using the ELK stack, which comprises ElasticSearch, Logstash, and Kibana. ElasticSearch is an open-source, scalable, real-time search engine primarily used for data organization and accessibility. Logstash is an open-source event management tool for centralizing and analyzing

structured and unstructured data. Kibana is a flexible web interface that allows customization to create complex charts, graphs, and visual representations. Figure 1.8 illustrates the modules that make up this framework.

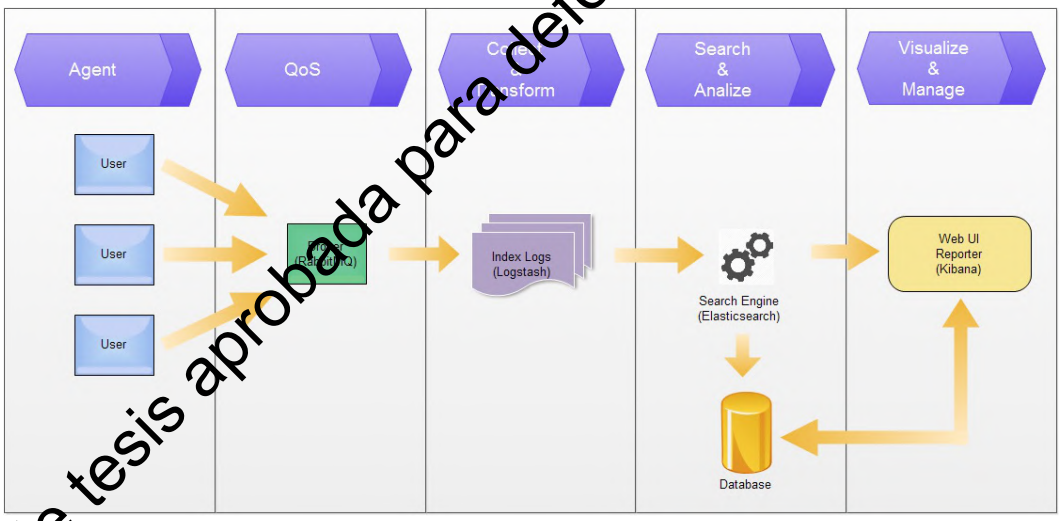


**Figure 1.8:** FraudFind Framework

FraudFind consists of extracting data from different sources of information through agents installed in workstations, which collect behavioral data and send it in an organized way, reporting its activity to the central server. All this is aimed at ensuring the security of the transactions generated by the users trying to identify possible acts of fraud by analyzing human behavior and treating the results.

When there are different sources of information, we find inconsistency in the logs, given that the formats are different. This represents a problem since administrators require access to this information for analysis, and there is difficulty in searching in different formats.

The possible violation of privacy is a factor that should be considered when implementing this solution within a company. Legal data protection regulations should be considered in a given region.

Considering human behavior factors, detecting unusual transactions that would not have been considered using traditional audit methods is possible. These behavior patterns can be found in the information that users generate when using the different applications on a workstation. The collected data is examined using data mining techniques to obtain patterns of suspicious behavior evidencing possible fraudulent behavior.

This contribution was published in the **Conference:** *IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), 2018.*

# REFERENCES


[1] ACFE Asociacion de Examinadores de Fraudes Certificados. (Date last accessed 15-July-2022).

[2] Abdul Khalique Shaikh and Amril Nazir. A novel dynamic approach to identifying suspicious customers in money transactions. *Int. J. Bus. Intell. Data Min.*, 17(2):143–158, 2020.

[3] Prabin Kumar Panigrahi. A framework for discovering internal financial fraud using analytics. In *2011 International Conference on Communication Systems and Network Technologies*, pages 323–327. IEEE, 2011.

[4] K Sayal and G Singh. What role does human behaviour play in corporate frauds. *Econ. Political Wkly*, 55, 2020.

[5] Dawn Cappelli, Andrew Moore, Randall Trzeciak, and Timothy J Shimeall. Common sense guide to prevention and detection of insider threats, 2009.

[6] Rasha Kassem. Detecting asset misappropriation: a framework for external auditors. *International Journal of Accounting, Auditing and Performance Evaluation (IJAAPE)*, 10(1), 2014.

[7] Thanasak Ruankaew. The fraud factors. *International Journal of Management and Administrative Sciences*, 2(2):1–5, 2013.

[8] Noorhayati Mansor and Rabiu Abdullahi. Fraud triangle theory and fraud diamond theory. understanding the convergent and divergent for future research. *International Journal of Academic Research in Accounting, Finance and Management Science*, 1(4):38–45, 2015.

[9] Gianluca Gabrielli and Alice Medioli. An overview of instruments and tools to detect fraudulent financial statements. *Univ. J. Account. Financ*, 7:76–82, 2019.

[10] Dragomir Dimitrijević and Zoran Kalinić. Software tools usage in fraud detection and prevention in governmental and external audit organizations in the republic of serbia1. *Knowledge–Economy–Society; Cracow University of Economics: Cracow, Poland*, page 71, 2017.

[11] Efstathios Kirkos, Charalambos Spathis, and Yannis Manolopoulos. Data mining techniques for the detection of fraudulent financial statements. *Expert systems with applications*, 32(4):995–1003, 2007.

[12] Meenatkshi R and Sivaranjani K. Fraud detection in financial statement using data mining technique and performance analysis. *JCTA*, 9:407–413, 2016.

[13] Khaled Gubran Al-Hashedi and Pritheega Magalingam. Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review*, 40:100402, 2021.

[14] Wenkai Deng, Ziming Huang, Jiachen Zhang, and Junyan Xu. A data mining based system for transaction fraud detection. In *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, pages 542–545. IEEE, 2021.

[15] Clifton Phua, Vincent Lee, Kate Smith, and Ross Gayler. A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*, 2010.

[16] Francisco Mota, Nuno Abreu, Tiago Guimarães, and Manuel Santos. A data mining study on pressure ulcers. pages 251–258, 01 2019.

[17] Oded Maimon and Lior Rokach. *The Data Mining and Knowledge Discovery Handbook*, volume 1. 01 2005.

[18] Nont Kanungsukkasem and Teerapong Leelanupab. Financial latent dirichlet allocation (finlda): Feature extraction in text and data mining for financial time series prediction. *IEEE Access*, PP:1–1, 05 2019.

[19] Reverso Context.

[20] Sentence Dict.

[21] Random Word Generator.

[22] Marco Sáncheza, Verónica Olmedo, Carlos Narvaeza, Myriam Hernándeza, and Luis Urquiza-Aguiar. Generation of a synthetic dataset for the study of fraud through deep learning techniques.

# 2 FRAUD DETECTION USING THE FRAUD TRIANGLE THEORY AND DATA MINING TECHNIQUES: A LITERATURE REVIEW

Marco Sánchez[1], Luis Urquiza[1], José Estrada[1]

[1]Departamento de Informática y Ciencias de la Computación, Escuela Politécnica Nacional, Ladrón de Guevara E11-253, Quito 170517, Ecuador

[2]Departamento de Electrónica, Telecomunicaciones y Redes de Información, Escuela Politécnica Nacional, Ladrón de Guevara E11-253, Quito 170517, Ecuador

## 2.1 ABSTRACT

Fraud entails deception to obtain illegal gains; thus, it is mainly evidenced within financial institutions and is a matter of general interest. The problem is particularly complex since fraud perpetrators could belong to any position, from top managers to payroll employees. Fraud detection has traditionally been performed by auditors, who mainly employ manual techniques. It could take too long to process fraud-related evidence. Data mining, machine learning, and, recently, deep learning strategies are being used to automate this type of processing. Many related techniques have been developed to analyze, detect, and prevent fraud-related behavior, with the fraud triangle associated with the classic auditing model being one of the most important of these. This work aims to review current work related to fraud detection that uses the fraud triangle and machine learning and deep learning techniques. We used the Kitchenham methodology to analyze the research works related to fraud detection from the last decade. This review provides evidence that fraud is an area of active investigation. Several works related to fraud detection using machine learning techniques were identified without the evidence that they incorporated the fraud triangle as a method for more efficient analysis.

**KEY WORDS:** fraud; machine learning; cybersecurity; human behavior.



## 2.2 INTRODUCTION

Fraud has recently increased considerably, affecting financial institutionsánd customersínterests. A study conducted by Price Waterhouse Coopers found that 30 % of the companies that they surveyed had already been victims of fraud. Moreover, 80 % of their fraud was committed within the companies' ranks, especially in administrative areas, such as accounting, operations, sales, and management level, without leaving aside the customer service dependencies [1]. Fraud-related activities, generally unknown within a company, determine a series of irregularities and illicit acts characterized by intentional deception committed by fraudsters. Most of the anomalies detected are due to the lack of internal control mechanisms, and in such situations, scammers commit fraud by exploiting the weaknesses [2].

Fraud is considered a subset of internal threats, such as corruption, misappropriation of assets, and fraudulent declarations, among others [3]. In a more formal definition, fraud is "the use of one's occupation for personal enrichment through the misuse or deliberate misapplication of the resources or assets of the employing organization," according to the Association of Certified Fraud Examiners (ACFE) [4]. The ability to commit this type of activity is based on the weakness of the control mechanisms that institutions and companies have. In such circumstances, fraudsters commit fraud by taking advantage of these weaknesses.

Since humans commit it, fraud is tightly coupled with human behavior. Thus, understanding the motivations of perpetrators or their psychological and personality traits that drive them to cross ethical boundaries can provide a new perspective for fraud detection [5]. Currently, there are different solutions [6] for detecting fraud, which are focused on the use of different tools that perform statistical and parametric analyses based on data mining techniques, as well as analyses of behavior, but none of them solve the problem of timely fraud detection [7].

Given the complexity of analyzing human behavior to detect fraud, some approaches in this line have been proposed to tackle some of the issues involved in this task. For instance, some works aimed to improve the precision and increase the speed of data processing through a hybrid automatic learning system [8] or through incremental learning [9]. Another challenge for fraud detection is the lack of data from which detection systems learn, and

[10] proposed a fraud-detection system that does not require previous fraudulent examples. However, even when the data are available, large and small datasets should be addressed differently [11]. As a human behavior, fraud detection is a multidimensional problem, and so are some of the fraud-detection mechanisms proposed in the literature [12, 13].

There is a consensus that prevention should be a priority to minimize fraud through proper risk management. Avoiding fraud saves time and financial resources since detecting it after it occurs means the stolen assets are practically irrecoverable. To enhance fraud prevention, organizations should focus on the root of the problem by identifying the causes that lead people to commit fraud and to understand their behavior [14]. Many theories have attempted to answer this question, and the most frequently cited in this context are Cressey's Fraud Triangle Theory (FTT) and Wolf and Hermanson's Diamond Fraud Theory (FDT) [15]. Both approaches analyze how perpetrators go so far as to commit fraud, which is discussed below.

The study of fraud and its analysis is best explained with the help of the Fraud Triangle Theory (FTT), proposed by Donald R. Cressey, a leading expert in the sociology of crime. Cressey investigated why people committed fraud and determined their responses based on pressure, opportunity, and rationalization. This theory also mentions that these elements occur consecutively to provoke the desire to commit fraud. The first necessary element is perceived pressure, which is related to the motivation and drive behind the fraudulent actions of an individual. This motivation often occurs in people under some form of financial stress [16]. The second element, perceived opportunity, is the action behind the crime and the ability to commit it. Finally, the third component, rationalization, concerns the idea that individuals can rationalize their dishonest acts, making their illegal actions seem justified and acceptable [17].

The FDT considered an extended version of the FTT, integrates a new vertex with the three that were already known—capacity [18]. Despite the cohesion among the three vertices of pressure, opportunity, and rationalization, it is unlikely that people will commit fraud unless they have the capacity (considered the fourth vertex). In other words, the potential perpetrator must have the skills and ability to commit fraud [19]. Various theories of fraud have been used to explain the motivation of this phenomenon. The FTT and FDT can be effectively used to detect the possibility of corporate fraud, where the measurement of all of the associated variables will depend to a great extent on the data used for the study, whether public or private [20].

Fraud analysis, when supported by data mining techniques helps reduce the manual parts of the detection/verification process and makes the search for fraud more efficient. It is impossible to guarantee people's proper moral and ethical behavior, especially in the workplace. Due to this reality, a valid option for identifying possible evidence of fraud from available data is to use automatic learning algorithms. Many works cover fraud detection and use data mining techniques as the primary focus [21, 22, 23, 24]. Two criticisms of data-mining-based fraud-detection research are frequently raised: the deficiency of the actual public data available in this domain for conducting experiments [25]—appropriate access to data for researching this area is extremely difficult due to privacy—and the lack of well-documented and published methods and techniques.



## 2.3 RELATED WORK

Here, we describe some systematic reviews whose main objectives were analyzing and detecting fraud using automatic learning techniques and the application of fraud theories.

Phua et al. [25] carried out a survey in which they identified the limitations of fraud-detection methods and techniques and showed that this field can benefit from other related areas. Specifically, unsupervised approaches may benefit from existing monitoring systems and text extraction, semi-supervised, and game-theoretical approaches; spam and intrusion detection communities can contribute to future fraud-detection investigations. However, above all, the authors focused on the nature of the information and excitedly reflected on the investigation of fraud detection based on data mining. They also referred to the scarcity of publicly available and real experimental data and the lack of well-documented and published methods and techniques.

Zhou et al. [26] concluded that most fraud-detection systems employ at least one supervised learning method and that unsupervised and semi-supervised learning methods are also used. The study showed that these techniques can be used alone or in combination to build more robust classifiers and that, without losing generality, these approaches are relatively successful in detecting fraud and credit scoring. They mentioned that fraud detection and data-mining-based credit scoring are subject to the same classification-related issues, such as feature engineering, parameter selection, and hyperparameter tuning. The authors also observed that fraud-related data are not abundant enough for investigators to train and test their models and that complex financial scenarios are nearly impossible to represent. They

explained that fraud detection must constantly evolve, particularly depending on the industry in which it is applied.

The authors of [27] performed a meta-analysis to establish the effect of mapping data samples from fraudulent companies to non-fraudulent companies using classification methods by comparing the general classification precision found in the literature. The results indicated that fraudulent samples could be matched equally to non-fraudulent samples (1:1 data mapping) or unevenly mapped using a one-to-many ratio to increase the sample size proportionally. Based on this meta-analysis, machine learning approaches can achieve better classification precision compared to statistical techniques, specifically when the availability of sample data is low. Furthermore, machine learning classification approaches can obtain high classification precision with a dataset with 1:1 mapping.

The results mentioned by the authors of [28] clearly show that data mining techniques have been widely applied for fraud detection in other fields, such as insurance, corporate, and credit card fraud. In this line, we found a lack of research on mortgage fraud, money laundering, and security fraud.

The main data mining techniques for financial fraud are logistical models that provide immediate solutions to the problems inherent in detecting and classifying fraudulent data. The authors of [29] conducted a review of the literature to address the following research questions related to financial statement fraud (FSF): (1) Can FSF be detected, how likely is it, and how can it be done? (2) What data characteristics can be used to predict FSF? (3) What kind of algorithm can be used to detect FSF? (4) How can detection performance be measured? (5) How effective are these algorithms in terms of detecting fraud? This work presents a generic framework to guide this analysis.

The reviews mentioned above have something in common: They try to unveil the main techniques used for fraud detection, such as machine learning methods (supervised, unsupervised, and semi-supervised), and try to identify which of these are more effective. This analysis was carried out in different scenarios, contrasting the results obtained and specifying the study area in which they are most accurate. We could not find studies linking fraud detection using machine learning techniques and the Fraud Triangle Theory.

Finally, we must comment on some theories to understand fraud detection. Studies such as [15] analyzed the convergence and divergence of two classic theories of fraud: the triangle theory and the diamond theory. The concept of fraud and the convergence of the two clas-

sical theories were examined there. This work also discussed the differentiation between them. In doing so, the similarities and differences between these theories were highlighted and appreciated. A discussion of the two approaches contributes to understanding fraud, especially for fraud professionals and examiners.



### 2.3.1 Contribution

This research aims to compile the literature related to fraud detection from two perspectives. On the one hand, we analyze works considering human behavior as an inherent risk factor in this problem, especially using the FTT and FDT. Beyond exploring these theories, on the other hand, our review analyzes different works where machine learning techniques have been used for fraud detection. Moreover, we look for works that integrate ML techniques with behavior-based fraud theories, such as the FTT and FDT. To do this, we used the well-known methodology of Barbara Kitchenham and formulated three research questions. As a result, we provide an up-to-date and comprehensive analysis of the subject. It will help identify, investigate, and evaluate the causes that lead to fraud and detect it. This study can guide further research on the topic in areas that the investigation has not considered. The rest of this paper is organized as follows. Section 2.4 addresses the methodology used to perform this review. Then, Section 2.5 summarizes our findings. After that, we discuss the weaknesses and strengths of the techniques identified in Section 2.6. Finally, Section 2.7 concludes and describes future work.

## 2.4 MATERIALS AND METHODS

A systematic literature review (SLR) was carried out for this research work. According to [25], the purpose of an SLR is to provide a complete list of all studies related to specific subject areas. Meanwhile, traditional reviews attempt to summarize the results of several studies. An SLR uses an evidence-based approach to meticulously search for relevant studies within a context to answer predefined research questions and select, evaluate, and critically analyze the findings to answer those research questions; this is done by following the recommendations reported in [30]. Considering the guidelines and recommendations described by Barbara Kitchenham [31], a systematic literature review must follow the methodological process illustrated in Figure 2.1.

### 2.4.1 Research Questions

As we stated, this article aims to review and summarize the works related to fraud detection that is performed by using machine learning techniques or the Fraud Triangle Theory. We do not restrict our search to any specific knowledge. The SLR research questions (RQs) that we intend to answer in this paper are the following:

1. RQ1: How can fraud be detected by analyzing human behavior by applying fraud theories?
2. RQ2: What machine or deep learning techniques are used to detect fraud?
3. RQ3: Using machine learning techniques, how can fraud cases be detected by analyzing human behavior associated with the Fraud Triangle Theory?



### 2.4.2 Keywords

We looked for scientific publications related to fraud detection, its process of identification, and its application to answering our research questions. We specifically targeted works focused on fraud that relied on machine learning techniques or the Fraud Triangle Theory. To this end, we created a base list of keywords built from the keywords found in related research, as shown in Table 2.1.

**Table 2.1:** Keywords.

| Title 1 | Title 2 | Title 3 |
|---|---|---|
| 1 | fraud | FR |
| 2 | fraud detection | FD |
| 3 | fraud triangle theory | FTT |
| 4 | fraud diamond theory | FDT |
| 5 | human behavior | HB |
| 6 | behavior patterns | BP |
| 7 | data mining | DT |
| 8 | machine learning | ML |
| 9 | deep learning | DL |

### 2.4.3 Search Strategy

We employed the guidelines from [32, 33] to define a search strategy to retrieve as many relevant documents as possible. Our search strategy is described below.

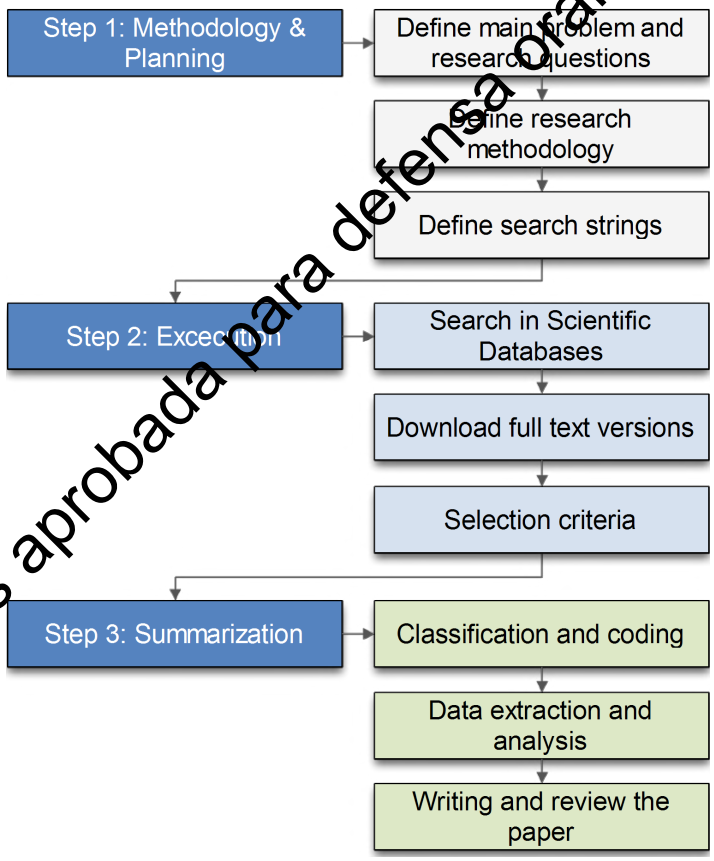




**Figure 2.1:** Methodology applied in the systematic literature review (SLR).

#### 2.4.3.1 Search Method

To find the most relevant publications for the topic addressed in this work, we queried the following databases: IEEEXplore, ScienceDirect, ACM Digital Library, and Scopus. We chose these databases because they offer the most essential and high-impact full-text journals and conference proceedings that cover the ML and FD fields in general. We carried out the searches in the titles, keywords, and abstracts of articles using the combinations of terms introduced in the following section.

#### 2.4.3.2 Search Terms

The search string was designed according to what was mentioned in [34]. Based on the research questions, we constructed the following relationships: (“Data mining” OR “Machine learning” OR “Deep Learning”) AND (“Detection Fraud” OR “Internal Fraud” OR “Fraud Triangle” OR “Diamond Triangle” OR “Human Behavior”). These search terms were combined using “AND” operators to build the search string. The search terms in the string only

**Table 2.2:** Inclusion/exclusion criteria.

| No | Inclusion Criteria |
|---|---|
| IC1 | Indexed publications not older than ten years. |
| IC2 | Scope of study: Computer Science |
| IC3 | Primary studies (journal or articles). |
| IC4 | Papers that discuss aspects regarding fraud detection. |
| IC5 | The investigations considered have information relevant to the research questions. |
| **No** | **Exclusion Criteria** |
| EC1 | Papers in which the language differs from English cannot be selected. |
| EC2 | Papers that are not available for reading and data collection (papers that are only accessible by paying or are not provided by the search engine) cannot be selected. |
| EC3 | Duplicated papers cannot be selected. |
| EC4 | Publications that do not meet any of the inclusion criteria cannot be selected. |
| EC5 | Publications that do not describe scientific methodology cannot be selected. |



matched the title, abstract, and keywords of the digital databases' articles. It is essential to find the correct search field or combination, be it the title, abstract, or full text, to apply in the search string and, thus, obtain effective results. In many cases, searching only by the "title" does not always provide the most relevant publications. Therefore, it can be necessary to include the "abstract" and, in other cases, "the complete document" of the related publications.

#### 2.4.3.3 Selection of Papers

Since the searches in the articles' full text resulted in many irrelevant publications, we decided to apply the search criteria by incorporating the "abstracts" of the papers. This means an article was selected as a potential candidate if its title or abstract contained the keywords defined in the search string. As a first filter, we evaluated each paper's title and abstract according to the inclusion and exclusion criteria (see Table 2.2). We selected the articles within the scope of the research questions. We thoroughly and entirely read the previously selected articles (which passed the first filter) as a second filter. The papers were included or excluded according to the inclusion and exclusion criteria. We will focus next on explaining the inclusion/exclusion criteria. Additionally, the search was limited to research written in English and published since 2010 [35].

### 2.4.4 Study Selection

As shown in Figure 2.2, the selection of studies was performed through the following processes [36]:

1. Identification: The keywords were selected from the databases listed above according to the research questions mentioned in the search method section. The search string was applied only to the title and abstract, as a full-text search would produce many irrelevant results [37]. The search period went from 2010 to 2021.
2. Filter: All possible primary studies' titles, abstracts, and keywords were checked against the inclusion and exclusion criteria. If it was difficult to determine whether an article should be included or not, it was reserved for the next phase.
3. Eligibility: At this stage, a complete reading of the text was done to determine if the article should be included according to the inclusion and exclusion criteria.
4. Data extraction: After filtering, data were extracted from the selected studies to answer RQ1–RQ3.



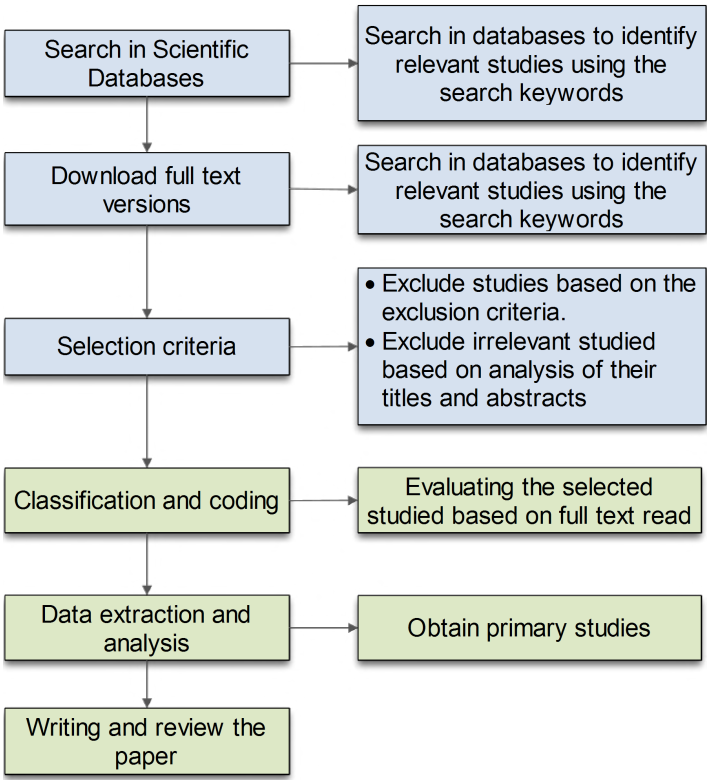


**Figure 2.2:** Process of the selection of studies.

### 2.4.5 Quality Assessment

We assessed their quality after selecting several primary studies based on the inclusion and exclusion criteria. Following the guidelines in [26], three quality assessment (QA) questions were defined to measure each proposal's research quality and provide a quantitative comparison between the research works considered. The criteria were based on three quality assessment (QA) questions:

1. Are the topics covered in the article relevant to fraud detection? Yes: It explicitly describes the topics related to fraud detection by applying ML techniques through the FTT. Partially: Only a few are mentioned. No: It neither describes nor mentions topics related to fraud detection using ML techniques through the FTT.

2. Were the limitations for the study of fraud detection detailed? Yes: It clearly explained the limitations related to fraud detection by applying ML techniques through the FTT. Partially: It mentioned the limitations but did not explain why. No: It did not mention the limitations.

3. Did the study address systematic research? Yes: The study was developed systematically and applied an adequate methodology to obtain reliable findings. Partially: The study was developed systematically and used a proper methodology but did not provide details. No: The study was not explained clearly, and the authors did not apply an adequate methodology.

The scoring procedure was defined as follows: Y (Yes = 1), P (Partially = 0.5), N (No = 0), or Unknown (i.e., the information was not specified).

### 2.4.6 Data Extraction and Analysis

This section describes the data extraction process performed with the selected papers and the analysis of the data extracted to answer the research questions of this SLR. We extracted the required data from previously selected works that were accordingly classified to answer the research questions, as shown in Table 2.3. The data extraction form used for all selected primary studies is indicated to conduct an in-depth analysis.

**Table 2.3:** Data extraction form

| No | Extracted Data | Description | Type |
|---|---|---|---|
| 1 | Identity of the study<br>Bibliographic references<br>Type of study | Unique identity for the study | General |
| 2 | Bibliographic references | Authors, year of publication, title, and source of publication | General |
| 3 | Type of study | Book, journal paper, conference paper, workshop paper | General |
| 4 | The theories employed | Description of the detection of fraud by applying the FTT and HB | RQ1 |
| 5 | The techniques considered | Description of the detection of fraud by applying ML/DM techniques | RQ2 |
| 6 | Combination of techniques and theories used | Description of the analysis of theories and techniques used to detect fraud | RQ3 |
| 7 | Findings and Contributions | Indication of the findings and contributions of the study | General |



We extracted the most representative papers related to the research questions based on the search string and associated terms. The results of the data analysis are presented in the next section.

### 2.4.7 Synthesis

Many papers could contain keywords used in the search string, but they could be irrelevant to our research questions. Therefore, a careful selection of documents should include only those containing helpful information concerning the research approach and the answers to the different research questions. As shown in Figure 2.3, we first searched each data source separately to join later the results obtained from the various sources of information, resulting in a total of 1891 papers. We obtained the most articles from Scopus, representing around 50 % of all documents.

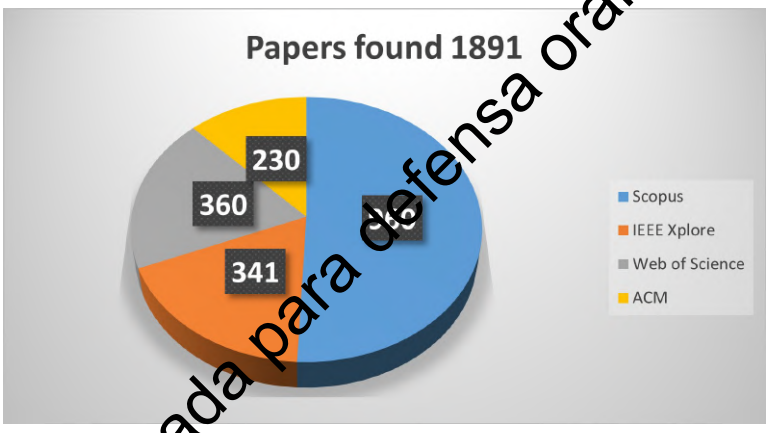


**Figura 2.3:** Studies retrieved through search engines.



Table 2.4 shows the number of articles found per source according to the search for keywords related to the search strings in the selected databases. The second column shows the results of the initial selection of papers found in each source. Below is the number of articles chosen after removing the exclusion criteria. The number of articles that were selected after eliminating duplicate articles is presented in the fourth column. Finally, the papers from each source selected after the inclusion process are presented.

**Table 2.4:** Number of papers found through the selection process.

| Source | Papers Found | Abstract and Title | Duplicity | Selected |
|---|---|---|---|---|
| Scopus | 960 | 77 | 48 | 16 |
| IEEE | 341 | 68 | 31 | 7 |
| WoC | 360 | 61 | 16 | 9 |
| ACM | 230 | 48 | 11 | 4 |
| Total | 1891 | 254 | 106 | 32 |

It was necessary to refine the papers obtained by previously eliminating irrelevant studies to ensure that the works complied with the established selection criteria. Our search in the databases, applying the search string to only the titles and abstracts of the articles, and selecting articles published during the last eleven years yielded 1891 records. After using the exclusion criteria on these records, we obtained 254 studies. The analysis of the duplicity of such studies enabled us to find 106 relevant papers for a full-text review. Finally, after a full-text assessment, 32 studies [38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69] were identified as a result of the analysis through the SLR technique. Therefore, a total of 32 publications met all of the inclusion criteria. The selection of studies from the initial search identification phase and the final number of included studies are presented in Figure 2.4. As initially proposed

and to ensure that the resulting reviews contained relevant information, we read the full text of the 32 studies to verify if they fit our adopted selection criteria. As a result, all of these publications represented our final set of primary studies.



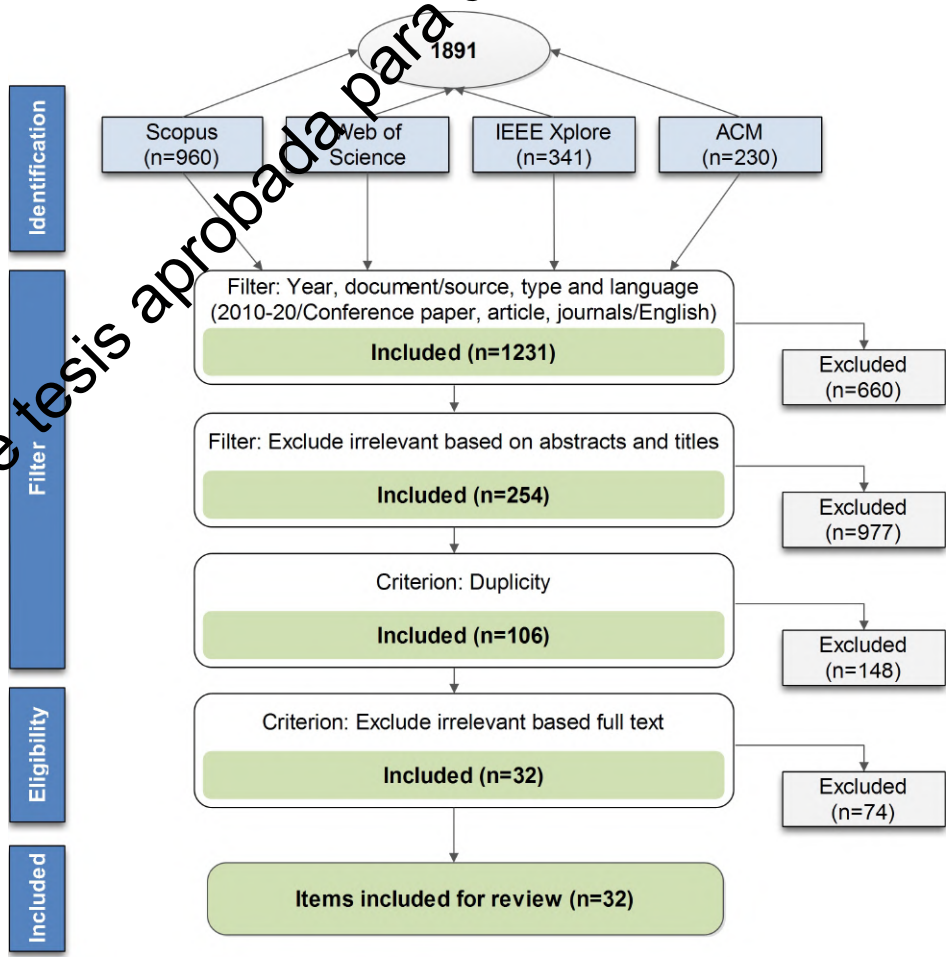


**Figure 2.4:** Steps followed to narrow the search results.

Regarding the types of publications where the selected papers were available, we found that 50 % of them had been published in conferences and 50 % in journals.

Table 2.5 shows the number of citations of the selected articles. The data presented (column cited) only approximates the citation rates and is not intended for comparisons among studies.

Regarding the selected articles'publication period, 32 studies were published between 2010 and 2021. Furthermore, as shown in Figure 2.5, 2010, 2015, 2016, and 2017 had the most significant articles, while 2011, 2012, 2019, and 2020 had the lowest numbers.

**Table 2.5:** Numbers of selected studies by type.

| # | Cited | # | Cited | # | Cited | # | Cited |
|---|---|---|---|---|---|---|---|
| [38] | 905 | [48] | 6 | [58] | 43 | [68] | 954 |
| [39] | 16 | [49] | 6 | [59] | 23 | [55] | 6 |
| [40] | 20 | [50] | 431 | [60] | 258 | | |
| [41] | 3 | [51] | 9 | [61] | 5 | | |
| [42] | 55 | [52] | 0 | [62] | 133 | | |
| [43] | 18 | [53] | 16 | [63] | 90 | | |
| [70] | 120 | [54] | 55 | [64] | 29 | | |
| [45] | 11 | [65] | 7 | [46] | 22 | | |
| [56] | 7 | [66] | 3 | [47] | 22 | | |
| [57] | 209 | [67] | 4 | [69] | 6 | | |

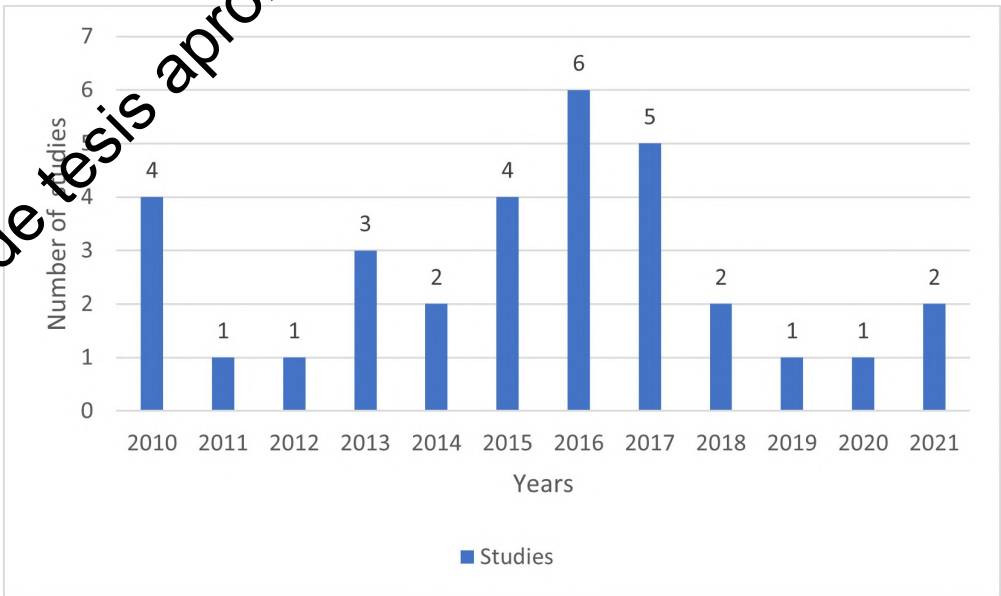


**Figure 2.5:** Number of articles by year of publication.



## 2.5 RESULTS

As a result of our methodology, we found 32 documents published between 2010 and 2021 that covered the most representative work on the topic of this paper. We focused only on peer-reviewed papers from journals and conferences. All of them were obtained from searching for fraud-related topics in four scientific libraries. Table 2.6 shows a matrix built using the topics most closely related to the research questions and with references to the corresponding articles. As can be seen, each column identifies a relevant topic associated with the research questions. We can see that seven works were found for RQ1(Fraud Detection + Human Behavior + Fraud theory). In contrast, for RQ2 (Fraud Detection + ML/DM techniques), 24 works were found, while for RQ3 (Fraud Detection + Human Behavior + ML/ DM + Fraud theory), only one study was found. So, there is room for improving fraud detection

because RQ3 combines most of the topics in the other research questions.

**Table 2.6:** Topics related to the research questions.

| # | Ref | Fraud Detection | Human Behavior | ML/DM Techniques | Fraud Theory |
|---|---|---|---|---|---|
| 1 | [38] | RQ1 | RQ1 | | RQ1 |
| 2 | [39] | RQ1 | RQ1 | | RQ1 |
| 3 | [40] | RQ1 | RQ1 | | RQ1 |
| 4 | [41] | RQ1 | RQ1 | | RQ1 |
| 5 | [42] | RQ1 | RQ1 | | RQ1 |
| 6 | [43] | RQ1 | RQ1 | | RQ1 |
| 7 | [70] | RQ1 | RQ1 | | RQ1 |
| 8 | [45] | RQ2 | | RQ2 | |
| 9 | [46] | RQ2 | | RQ2 | |
| 10 | [47] | RQ2 | | RQ2 | |
| 11 | [48] | RQ2 | | RQ2 | |
| 12 | [49] | RQ2 | | RQ2 | |
| 13 | [50] | RQ2 | | RQ2 | |
| 14 | [51] | RQ2 | | RQ2 | |
| 15 | [52] | RQ2 | | RQ2 | |
| 16 | [53] | RQ2 | | RQ2 | |
| 17 | [54] | RQ2 | | RQ2 | |
| 18 | [55] | RQ2 | | RQ2 | |
| 19 | [56] | RQ2 | | RQ2 | |
| 20 | [57] | RQ2 | | RQ2 | |
| 21 | [58] | RQ2 | | RQ2 | |
| 22 | [59] | RQ2 | | RQ2 | |
| 23 | [60] | RQ2 | | RQ2 | |
| 24 | [61] | RQ2 | | RQ2 | |
| 25 | [62] | RQ2 | | RQ2 | |
| 26 | [63] | RQ2 | | RQ2 | |
| 27 | [64] | RQ2 | | RQ2 | |
| 28 | [65] | RQ2 | | RQ2 | |
| 29 | [66] | RQ2 | | RQ2 | |
| 30 | [67] | RQ2 | | RQ2 | |
| 31 | [68] | RQ2 | | RQ2 | |
| 32 | [69] | RQ3 | RQ3 | RQ3 | RQ3 |

Table 2.7 shows the works found vs. the research question frequencies. As can be seen, RQ2 is the most frequently investigated. It accounts for 75 %. Only one paper was found for RQ3, accounting for 3.13 %, and RQ1 accounts for 21.88 %.

**Table 2.7:** Frequencies of the works found.

| RQ | Study Identifier | Frequency | Percentage |
|---|---|---|---|
| 1 | [38, 39, 40, 41, 42, 43, 70] | 7 | 21.88 |
| 2 | [45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56] [57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68] | 24 | 75 |
| 3 | [69] | 1 | 3.13 |

### 2.5.1 RQ1: How Can Fraud Be Detected by Analyzing Human Behavior by Applying Fraud Theories?

This section details the results obtained from the analysis of research papers that relate fraud detection with the point of view of human behavior by applying the Fraud Triangle Theory. The investigation is intended to answer RQ1. We answer this question by analyzing the number of documents linked to the research question. According to Table 2.6, seven works were found. Hoyer et al. [38] proposed a prototype in a generic architectural model that considers the factors of the fraud triangle. In this way, in addition to the analysis applied as part of a traditional fraud audit, human behavior is considered. By doing this, the transactions examined by an auditor can be better differentiated and prioritized. Behavioral patterns are found through the incorporation of the human factor. These patterns appear in multiple sources of information, especially in users' data, such as in e-mails, messages, network traffic, and system records from which evidence of fraud can be extracted.

Sanchez et al. [39] presented a framework that identifies people who commit fraud and is supported by the Fraud Triangle Theory. This proposal is based on the use of a continuous audit that is installed on user devices, collects information from agents, and employs the collection of phrases. They are subsequently analyzed to identify fraud patterns through analyzing human behavior and the treatment of the results. In [40], based on primary data on the behavior of perpetrators who commit fraud, the authors showed the complementarity between an ex-post analysis and the existing literature on this topic. They suggested that the presence or absence of fraudulent intent can be assessed by scrutinizing human behavior. Mackevicius and Giriunas [41] analyzed the Fraud Triangle Theory and presented its associated elements: "motives, possibilities, pressure, rationalization, incentive, and others." They offered a theoretical analysis of the fraud scales and their elements: motives, conditions, possibilities, and performance. To this end, the authors analyzed 265 respondents—including accountants, stakeholders, public officials, and inspectors in Central Java, Indonesia—by using structural equation modeling (SEM) with the AMOS analysis tools. In [42], the authors assessed the Fraud Triangle Theory and human behavior to study the factors of opportunity, financial processes, and rationalization. The authors emphasized the importance of psychological and moral aspects. The International Auditing Standard AI240 focuses on the auditor's responsibility to assess fraud in an audit of financial statements. The authors of [43] explored if the standard has been used effectively in Indonesia based on the

proposed fraud indicators through a fraud analysis. A questionnaire survey was conducted with three groups of auditors: external, internal, and government auditors. This study examined auditors' perceptions of the importance and existence of warning signs of financial fraud by using the fraud diamond. The findings indicate that the auditors could identify these red flags by giving them high scores. On the contrary, the scores were low regarding the "level of use."

Mekonnen et al. [70] presented an insider threat prevention and prediction model based on the fraud diamond by combining various approaches, techniques, and IT tools, as well as criminology and psychology. The deployment of this model involved collecting information about possible intentions by using privileged information within a context of preserving privacy, thus enabling high-risk insider threats to be identified while balancing privacy concerns.



### 2.5.2 RQ2: What Machine or Deep Learning Techniques Are Used to Detect Fraud?

his section reports the results of works that described the implementation of machine learning and data analysis for fraud detection. We aimed to identify this realm's most commonly used machine or deep learning techniques. Table 2.7 shows that this research question had the highest related works. Table 2.8 presents the articles'main focus, the ML/DL techniques used, and the dataset information. All of these articles are summarized below.

Some works enhance traditional security approaches. In [60], the need to use Process Information Systems (PAIS) software in organizations and the importance of fraud detection was investigated. They claimed this tool is necessary for organizations, as its flexibility raises fraud detection. The authors of [63] sought to design an artifact (hardware) for detecting communications from disgruntled employees through automated text-mining techniques. The artifact they developed extended the layered approach to combat internal security risks. They claimed that this phenomenon can be detected in e-mail repositories by using employee dissatisfaction as the primary indicator of fraud risk. Considering the methods of fraud detection based on simple comparisons, detection of associations, clustering, perdition, and outliers, an automated fraud-detection framework was proposed in [47]. The framework allowed fraud identification by using intelligent agents, data fusion techniques, and various data

mining techniques. In [67], the authors proposed the detection of bank fraud through data extraction techniques, association, grouping, forecasting, and classification to analyze customer data to identify patterns leading to fraud. To conclude this group of papers, West et al. suggested that a higher level of verification/authentication can be added to banking processes by identifying patterns. To do this, the authors reviewed key performance metrics used to detect financial fraud, focusing on credit card fraud. They compared the effectiveness of these metrics to detect if fraud was carried out. In addition, the performance of the application of various computational intelligence techniques to this problem's domain was also investigated, and the efficacy of different binary classification methods was explored.



**Table 2.8:** Summary of works that used machine or deep learning techniques to detect fraud.

| Ref. | Techniques [a] | Dataset | Main Focus |
|---|---|---|---|
| [45] | NN, DT, BN | N/A | Summarized and compared different datasets and algorithms for automated accounting fraud detection. |
| [46] | RF | Financial and non-financial data | Presented a hybrid detection model using machine learning and text mining methods for detecting financial fraud. |
| [47] | KDD | N/A | Automated fraud detection framework that allows fraud identification using intelligent agents, data fusion techniques, and data mining techniques. |
| [48] | KM | UCI Machine Learning Repository [71] | Modified k-means clustering algorithm for detecting outliers and removing them from the dataset to improve grouping precision. |
| [49] | C.45, KM, SVM, NB, CART | N/A | Categorized the different types of fraud and explained the best available data mining techniques. |
| [50] | NN | N/A | Used neural networks to correlate information from a variety of technologies and database sources to identify suspicious account activity. |

In [45], the authors summarized and compared different datasets and algorithms for automated accounting fraud detection. The selected works addressed mining algorithms, including

**Table 2.8:** *Summary of works that used machine or deep learning techniques to detect fraud. (Cont.)*

| Ref. | Techniques [a] | Dataset | Main Focus |
|---|---|---|---|
| [51] | KM Clustering and AdaBoost Classifier | Worldline and the Université Libre de Bruxelles | Presented a study on the use of clustering and classifier techniques and compared their precision for fraud detection. |
| [52] | SVM, ANN | Indonesian stock exchange (IDX) | Through the application of data mining algorithms, such SVM and ANN, the essential indicators for detecting financial fraud are profitability and efficiency. |
| [53] | MLR, SVM, and BN | N/A | Development of three multiple-class classifiers—MLR, SVM, and BN—as well as predictive tools for detecting and classifying misstatements according to the presence of intent of fraud. |
| [54] | MLFF, SVM, GP, GMDH, LR, PNN | N/A | Used data mining techniques that were tested on a dataset involving 202 Chinese companies and compared them with and without the selection of functions. |
| [55] | MLR, SVM, NN, ensemble techniques, and LDA | 10-K financial reports of documents (EDGAR) | For fraud detection in financial reporting, various techniques of natural language processing, and supervised machine learning are applied. |
| [56] | ANN | [72] | Identified a person of interest from a published corpus of Enron email data for research. |
| [57] | LR, NN, SVM, BN, DT, AdaBoost, and LogitBoost | [71] | Method based on Grammatical Genetic Programming (GBGP) through multi-objective optimization and set learning. They compared the proposed method with LR, NN, SVM, BN, DT, AdaBoost, and LogitBoost on four FFD datasets. |
| [58] | LR, ANN, KNN, SVM, Decision Stem, M5P Tree, J48 Tree, RF, and Decision Table | N/A | Explored the use of data mining methods to detect electronic ledger fraud through financial statements. |
| [59] | DRL | N/A | Applied DRL theory through two applications in banking and discussed its implementation for fraud detection. |
| [60] | Petri-Net, Heuristic | N/A | Used the Process Information Systems (PAIS) software in organizations for fraud detection. |
| [61] | DT, NB | N/A | Credit card fraud detection using supervised learning algorithms. |
| [62] | Luhn's and Hunt's | N/A | System that detects fraud in the processing of credit card transactions. |



statistical tests, regression analysis, NN, DT, BN, stack variables, etc. Regression analysis was widely used to hide data. Generally, the effect of detection and the precision of NN were

**Table 2.8:** *Summary of works that used machine or deep learning techniques to detect fraud. (Cont.)*

| Ref. | Techniques [a] | Dataset | Main Focus |
|---|---|---|---|
| [63] | NB | Email data | Designed an artifact (hardware) for detecting communications from disgruntled employees using automated text mining techniques. |
| [64] | MLCC | International financial service provider | Analyzed the use of a data mining approach to reduce the risk of internal fraud. |
| [65] | CNN, SLSTM, hybrid of CNN–LSTM. | Card transactions from an Indonesian bank | Explored three deep learning models for recognizing fraudulent card transactions. |
| [66] | DT, RF, NB | Twitter, and Facebook | Implementation of the document grouping algorithm as a set of classification algorithms and appropriate industry use cases. |
| [67] | Association, clustering, forecasting, and classification | N/A | Detection of bank fraud through data mining techniques. |
| [68] | GP, NN, SVM | UCSD-FICO | Key performance metrics used for Financial Fraud Detection (FFD) focusing on credit card fraud. |

[a] Neural Networks: NN; Decision Trees: DT; Bayesian Networks: BN; Random Forest: RF; K-means: KM; Support Vector Machine: SVM; Artificial Neural Network: ANN; Multinomial Logistic Regression: MLR; Multilayer Direct Feed Neural Network: MLFF; Genetic Programming: GP; Group Method of Data Management: GMDH; Logistic Regression: LR; Probabilistic NN: PNN; Binomial Logistic Regression: BLR; Latent Dirichlet Assignment: LDA; K-Nearest Neighbor: KNN; Deep Reinforcement Learning: DRL; Multivariate Latent Class Clustering: MLCC; Convolutional Neural Network: CNN; Stacked Long Short-Term Memory: SLSTM; Naive Bayes: NB.



higher than those of regression models. The overall conclusion was that pattern detection is better than detection by an unaided auditor. Due to the small size of the fraud samples, some publications reached decisions based on training samples and may have overestimated the effects of the models. In [46], S. Wang presented a hybrid detection model using machine learning and text mining methods for detecting financial fraud. This model used financial and non-financial data and employed two ways of selecting easy-to-explain characteristics. During the investigation, the author chose 120 fraudulent financial statements disclosed by the China Securities Regulatory Commission (CSRC) between 2007 and 2016. He compared the performance of five machine learning methods and found that the Random Forest method had the following advantages: (1) It is suitable for processing high-dimensional data; (2) it avoids overfitting to some extent; (3) it is robust and stable. Ravisankar et al. proposed using data mining techniques to identify companies that resort to financial statement fraud [54]. Specifically, the authors tested the MLFF, SVM, GP, GMDH, LR, and PNN techniques. The evaluation considered the role of feature selection and relied on a dataset involving 202

Chinese companies. Their results indicated that the PNN outperformed all methods without feature selection, and the GP and PNN outperformed others with feature selection and marginally equal precisions.

For other works that compared different ML methods, we found the following. In [53], the authors developed three multiple-class classifiers (MLR, SVM, and BN) to detect and classify misstatements according to the presence of fraud intent. Using the MetaCost tool, the authors conducted cost-sensitive learning and solved class imbalance and asymmetric misclassification costs. In [58], the use of data mining methods to detect fraud in electronic ledgers through financial statements was explored. The training techniques were used for the Linear Regression, ANN, KNN, SVM, Decision Stem, M5P Tree, J48 Tree, RF, and Decision Table. The authors of [61] detected credit card fraud using supervised learning algorithms, such as a DT and NB. Focusing on using or comparing ANNs with other methods, Vimal Kumar et al. [49] analyzed the challenges of detecting and preventing fraud in the banking industry when having insider information. The authors reviewed some data analysis techniques for detecting insider trading scams. Their work lists the best data mining techniques available (NN, DT, and Bayesian Belief Networks), which have been proposed by many researchers and employed in different industries. They concluded that the banking industry's primary requirements are fraud detection and prevention and that data mining techniques can help reduce fraud cases. In addition, the work in [50] proposed using NN to correlate information from various technological sources and databases to identify suspicious account activity. The work in [52] applied data mining algorithms, such as SVM and ANNs, to detect financial fraud. The authors stated that the essential financial fraud indicators are profitability and efficiency. Incorporating these factors improved the accuracy of the SVM algorithm to 88.37 %. The ANNs produced the highest precision, 90.97 %, for data without feature selection. In [56], Mohanty et al. aimed to identify a person of interest from the corpus of Enron email data released for research. They tried to detect fraudulent activities using an ANN with the Adam optimizer and ReLU activation functions. Their work achieved high precision regarding recall, accuracy, and F1 score.

Regarding unsupervised approaches, a proposal to detect outliers using a modified K-Means Clustering algorithm was presented in [48]. For this work, the detected outliers were removed from the dataset to improve the grouping precision. They also validated their approach against existing techniques and benchmark performance. The authors of [51] presented a study on using K-Means Clustering and the AdaBoost Classifier, comparing their

accuracies and performances with an analysis of the past and present models used for fraud detection. Regarding the use of more sophisticated techniques for the problem of fraud detection in financial reporting, the authors of [55] applied various natural language processing techniques and supervised machine learning, including BLR, SVM, NN, ensemble techniques, and LDA. They applied Latent Dirichlet Allocation (LDA) to a collection of 10-K financial reports of documents available in the EDGAR database of the United States Security and Exchange Commission to generate a frequency matrix of documents and topics. In addition, they applied evaluation metrics, such as the accuracy, receiver performance characteristic curve, and area under the curve, to evaluate the performance of each algorithm. To resolve problems for FFD, Li, and Wong, [57] proposed a new method based on GBGP through multi-objective optimization and set learning. They compared the proposed method with LR, NN, SVM, BN, DT, AdaBoost, bagging, and LogitBoost in four FFD datasets. The results showed the efficacy of the new approach on the given FFD problems, including two real-life situations. The authors of [59] applied the theory of DRL through two applications in banking and discussed its implementation for fraud detection. Using a DT with a combination of the Luhn algorithm and the Hunt algorithm, Save et al. [62] proposed a system that detects fraud in the processing of credit card transactions. The validation of the card number is done through the Luhn algorithm. The authors of [64] focused on detecting external fraud. Using a data mining approach to reduce the risk of internal fraud was also discussed. Consequently, a descriptive data mining strategy was applied instead of the widely used prediction data mining techniques. The authors employed a multivariate latent class clustering algorithm for a case firm's procurement data. Their results suggested that their technique helps assess the current internal fraud risk.

Exploring a deep learning model to learn short and long-term patterns from an unbalanced input dataset was an objective set by [65]. The data obtained were transactions of an Indonesian bank in 2016–2017 with binary labels (no fraud or fraud). They also explored the effects of sample ratios of non-fraud to fraud from 1 to 4 and three models: a convolutional neural network (CNN), short-term/long-term stacked memory (SLSTM), and a CNN–LSTM hybrid. Using the area under the ROC curve (AUC) as the model performance metric, CNN achieved the highest AUC for R = 1, 2, 3, 4, followed by the SLSTM and CNN–LSTM. The authors of [66] proposed implementing both the document clustering algorithm and a set of classification algorithms (DT, RF, and NB), along with industry-appropriate use cases. In addition, the performance of three classification algorithms was compared by calculating

the “Confusion Matrix,” which, in turn, helped us calculate performance measures such as “accuracy,” “precision,” and “recovery.”



### 2.5.3 RQ3: Using Machine Learning Techniques, How Can Fraud Cases Be Detected by Analyzing Human Behavior Associated with the Fraud Triangle Theory?

We found only one work related to this research question. This means we obtained few results when we tried keywords related to the topics most relevant to the research questions (Fraud Detection + Human Behavior + Machine Learning Techniques + Fraud Triangle Theory). Therefore, combining ML techniques and fraud-related theories needs further investigation because it would integrate two knowledge fields (psychology and data science) to improve fraud detection. In [69], the authors examined the aspects of the fraud triangle using data mining techniques to evaluate attributes such as pressure/incentive, opportunity, and attitude/rationalization, and, through the use of expert questionnaires, they discussed whether their suggestion agreed with the results obtained with the adoption of those techniques. The data extraction methods used in this research included logistic regression, decision trees (CART), and artificial neural networks (ANNs). They also compared data mining techniques and expert judgments. The ANNs and CART achieved training samples of 91.2 % (ANN) and 90.4 % (CART). They were tested with correct classification rates of 92.8 % (ANN) and 90.3 % (CART), which were more precise than those of logistic models, which only reached 83.7 % and 88.5 % of correct classification in the assessment of the presence of fraud.

### 2.5.4 Quality Assessment

Once the QA questions were defined, we evaluated the primary studies identified in the SLR. The score assigned to each study for each question is shown in Table 2.9.

The total of the accumulated scores from the QA questions can be observed in the “Total Score” row, showing that QA3 has 22 points, corresponding to 44.9 %, demonstrating that this question was more representative in the review. QA2 followed this with 33.68 %, and QA1 followed with 21.42 %. On the other hand, the last row identifies the percentage of points collected by the values assigned for a given QA question concerning the points ob-

**Table 2.9:** Quality assessment.

| # | QA-1 | QA-2 | QA-3 | Total Score | Max S |
|---|---|---|---|---|---|
| [38] | P | P | Y | 2 | 66.67 |
| [39] | P | P | Y | 2 | 66.67 |
| [40] | N | N | N | 0 | 0 |
| [41] | P | Y | N | 2 | 66.67 |
| [42] | N | N | N | 0 | 0 |
| [43] | N | N | N | 0 | 0 |
| [70] | P | P | Y | 2 | 66.67 |
| [45] | P | Y | Y | 2.5 | 83.33 |
| [46] | P | Y | Y | 2.5 | 83.33 |
| [47] | N | N | N | 0 | 0 |
| [48] | P | P | Y | 2 | 66.67 |
| [49] | P | Y | Y | 2.5 | 83.33 |
| [50] | P | P | Y | 2 | 66.67 |
| [51] | P | P | Y | 2 | 66.67 |
| [52] | P | P | Y | 2 | 66.67 |
| [53] | P | P | Y | 2 | 66.67 |
| [54] | N | N | N | 0 | 0 |
| [55] | P | P | Y | 2 | 66.67 |
| [56] | P | Y | Y | 2.5 | 83.33 |
| [57] | P | Y | Y | 2.5 | 83.33 |
| [58] | N | N | N | 0 | 0 |
| [59] | P | P | Y | 2 | 66.67 |
| [60] | P | Y | Y | 2.5 | 83.33 |
| [61] | N | N | N | 0 | 0 |
| [62] | N | N | N | 0 | 0 |
| [63] | P | Y | Y | 2.5 | 83.33 |
| [64] | 0 | 0 | 0 | 0 | 0 |
| [65] | P | P | Y | 2 | 66.67 |
| [66] | N | N | N | 0 | 0 |
| [67] | P | Y | Y | 2.5 | 83.33 |
| [68] | P | Y | Y | 2.5 | 83.33 |
| [69] | P | Y | Y | 2.5 | 83.33 |
| Total | 10.5 | 16.5 | 22 | 49 | |
| Max QA | 21.42 | 33.68 | 44.9 | 100 | |
| Total Score | 47.62 | 73.81 | 100 | | |



tained if each selected study received the highest score. Refs. [45, 46, 49, 56, 57, 60, 63, 67, 69] obtained the highest score of 2.5, which represents 83.33 % of the maximum score that a preliminary study could obtain; on the other hand, Refs. [38, 39, 41, 44, 48, 50, 51, 52, 53, 55, 59, 65] obtained a score of 2, that represents 66.67 % of the maximum score. Refs. [40, 42, 43, 47, 54, 58, 61, 62, 64, 66] failed to get any scores, which means that their title and abstract showed that they could answer the research question for this SLR, but after reviewing the full articles, no features related to fraud detection using machine learning techniques were discussed.

## 2.6 DISCUSSION



In this work, we have reviewed contributions related to fraud detection, with a particular emphasis on those addressing fraud detection from the perspective of modeling human behavior.

Applying techniques related to the analysis of human behavior allowed us to consider behavioral factors that could empower the detection of unusual transactions that would not have been considered if using traditional auditing methods. By observing people's behavior, it can be seen that the human factor is closely related to the Fraud Triangle Theory.

On the other hand, the use of machine learning techniques to detect fraud was also implemented in several works to predict behaviors related to this phenomenon. As a result of our research, many articles (24) addressed this approach. In this context, we found that mainly supervised and unsupervised algorithms are used for fraud detection analysis. The supervised strategy enables blocking fraud attempts based on fraudulent and non-fraudulent samples. This is used in rule-based detection, which automatically infers discriminatory rules from a labeled training set. In addition, regarding fraud detection, our research unveiled that supervised algorithms regularly have to deal with unbalanced classes, which might result in poor detection. Furthermore, these techniques are unable to identify new fraud patterns. Unsupervised learning, however, concentrates on discovering suspicious behavior as a proxy of fraud detection and, thus, does not require prior knowledge about verified fraudulent cases.

Our review focuses on fraud detection performed through machine learning techniques or analysis of human behavior based on the Fraud Triangle Theory. We tried to unveil how both approaches are addressed in the literature and how they may be jointly applied by answering three research questions.

By answering RQ1, keywords such as human behavior and theories related to fraud were linked, resulting in several related studies. The answer to RQ2 linked machine learning techniques with fraud detection; this question was the one that generated the most results. The analyzed questions produced results in a specific field. However, when trying to combine these fields by answering RQ3, we did not find works linking fraud detection using machine learning techniques with any theory related to fraud.

Despite the existence of works about detecting fraud in the areas of data mining and fraud theories, no literature reviews that jointly covered these two areas were identified. Table 2.10

presents a comparative summary of seven relevant SLRs and surveys performed in the area of fraud detection, including our contribution.

In the ”Context” column of Table 2.10, there are four SLRs that are exclusively related to some aspect of data mining [25, 26, 28, 29], while only one is related to some aspect of fraud theory [73], in addition to other approaches [74, 75]. The last row of Table 2.10 also presents information about the SLR covered in this document, the context of which explores data mining and fraud theories together, unlike the other seven presented in this table.



**Table 2.10:** Comparison of related systematic literature reviews.

| SLR Work | Year | Context | Period | Data Sources | # of Screened Works/ Primary Studies | Quality Assessment of Primary Studies |
|---|---|---|---|---|---|---|
| [25] | 2010 | Data-mining-based fraud detection | 2000–2010 | N/A | N/A | No evaluation criteria applied |
| [74] | 2020 | Fraud-detection metrics in business processes | N/A | 1, 4, 5, 7, 9, 14 | 12,000/75 | No well-defined evaluation criteria applied |
| [26] | 2018 | Data-mining-based fraud detection and credit scoring | N/A | N/A | N/A | No evaluation criteria applied |
| [75] | 2020 | Graph-based anomaly-detection approaches | 2007–2018 | 1, 2, 5, 9 | 585/39 | No evaluation criteria applied |
| [73] | 2019 | Fraud Triangle Theory | No specific | 7 | 1169/33 | Based on evaluation criteria proposed by authors |
| [28] | 2011 | Data mining techniques in financial fraud detection | 1997–2008 | 1, 2, 5, 9, 11, 12, 13 | 1200/49 | No well-defined evaluation criteria applied |
| [29] | 2007 | Data-mining-based financial fraud detection | N/A | N/A | N/A | No evaluation criteria applied |
| This SLR | 2021 | Fraud detection using the Fraud Triangle Theory and data mining techniques | 2010–2021 | 1, 2, 4, 10 | 1891/32 | Based on evaluation criteria proposed by [76] |

1: IEEE Xplore; 2: ACM DL; 3: Engineering Village (Compendex); 4: ISI Web of Science; 5: ScienceDirect; 6: Wiley Inter Science Journal; 7: Google Scholar; 8: Citeseer; 9:Springerlink; 10: Scopus; 11: Business Source Premier (EBSCO); 12: Emerald Full Text; 13: World Scientific Net; 14: ProQuest.

These SLRs were published between 2007 and 2020, with the novelty that some of them [74, 26, 29] do not mention the related search period. The research periods of [25, 28, 75] range from 10 to 11 years but include primary studies without making cuts in any specific year. Some works do not specify the sources of data, and those doing so report a variable number of data sources. Studies that mention data sources do not clearly explain their reasons for selecting them. On the other hand, four data sources were chosen for our research to maximize the probability of identifying relevant candidate works as primary studies.

The number of candidate articles from the data sources and the number of selected primary studies are presented in this table for each SLR. The differences in these numbers may be related to the context of each investigation, e.g., data sources used, keywords, etc. For our SLR, the number of reviewed works resulted from the searches in the different data sources combined with the chosen keywords. In contrast, the final number of primary studies was similar to those of other works. It should be noted that there are works that do not mention this metric.

Although quality evaluation is not a mandatory parameter in the structure of an SLR, according to [76], it is an essential contribution in this type of work to improve its quality. None of the analyzed works clearly showed how an evaluation was carried out. No criteria were mentioned for assessing the quality of the primary studies. Our work was based on the evaluation criteria proposed by [77].

## 2.7 CONCLUSIONS AND FUTURE WORK

Fraud detection is complex, as it requires the interpretation of human behavior, but this is not the only issue. The lack of data available for training or testing detection models significantly complicates the assessment of detection strategies. Even when data are available, unbalanced datasets are the norm in this domain.

Accordingly, very different approaches tackle the problem of fraud detection, as well as systematic literature reviews intended to address these limitations from a more global perspective. Thus, this research aimed to identify publications related to fraud detection using ML techniques based on the Fraud Triangle Theory. The proposed reference frameworks focus on developing tools that allow auditors to perform fraud analyses more efficiently by shortening their detection time through support from data mining techniques. Most of the works

concentrate on carrying out their analyses after fraud has been carried out to shorten the time taken to find results; thus, these proposals are reactive to such events.

Through this research, it was found that a significant number of research projects are being carried out in this specific area of fraud detection; in general, they have a solid level of maturity. The large number of publications in conferences and journals—representing 50 % and 50 % of primary studies, respectively—is substantial proof. In addition, the results of the quality evaluation carried out for the primary studies showed that the evaluation of their proposals was satisfactory in terms of the criteria of "relevance," "limitations," and "methodology." When we assumed an approach to fraud detection through data mining techniques and using fraud theories associated with human behavior, this SLR reveals very little evidence from studies supporting this approach since only one primary study was found, corresponding to 3.13 % of the studies. When we allowed partial coverage, that is, fraud detection by applying only data mining techniques, 24 primary studies (corresponding to 75 %) could be classified. On the other hand, when we analyzed the approach to the analysis and detection of fraud in which only theories related to fraud associated with human behavior were considered, seven primary studies (corresponding to 21.88 %) supported this approach.

In this sense, only one study with evidence of the use of data mining techniques, the application of fraud theories, and a corresponding analysis of human behavior to detect fraud was identified, which means there is a gap, and this is an appropriate field to investigate.

As future work, it is proposed that a review focused on detecting fraud and incorporating an analysis of the availability of data and the lack of access to this resource, including other data sources as possible alternatives, should be carried out.

# REFERENCES


[1] Abdul Khalique Shaikh and Amril Nazir. A novel dynamic approach to identifying suspicious customers in money transactions. *International Journal of Business Intelligence and Data Mining*, 17(2):143–158, 2020.

[2] Prabin Kumar Panigrahi. A framework for discovering internal financial fraud using analytics. In *2011 International Conference on Communication Systems and Network Technologies*, pages 323–327. IEEE, 2011.

[3] George Silowash, Dawn Cappelli, Andrew Moore, Randall Trzeciak, Timothy Shimeall, and Lori Flynn. Common sense guide to prevention and detection of insider threats 4th edition. *Carnegie Mellon University CyLab*, 12 2012.

[4] R Kassem. Detecting asset misappropriation: A framework for external auditors. *International Journal of Accounting Auditing and Performance Evaluation*, 10:1–42, 2014.

[5] Kangan Sayal and Gurparkash Singh. What role does human behaviour play in corporate frauds? *Economic and political weekly*, 55, 06 2020.

[6] Gianluca Gabrielli and Alice Medioli. An overview of instruments and tools to detect fraudulent financial statements. *Universal Journal of Accounting and Finance*, 7(3):76–82, 2019.

[7] Dragomir Dimitrijević and Zoran Kalinić. Software tools usage in fraud detection and prevention in governmental and external audit organizations in the republic of serbia1. *KNOWLEDGE–ECONOMY–SOCIETY*, page 71.

[8] Olena Vynokurova, Dmytro Peleshko, Oleksandr Bondarenko, Vadim Ilyasov, Vladislav Serzhantov, and Marta Peleshko. Hybrid machine learning system for solving fraud detection tasks. pages 1–5, 08 2020.

[9] Bertrand Lebichot, Gian Marco Paldino, Gianluca Bontempi, Wissam Siblini, Liyun He, and Frederic Oble. Incremental learning strategies for credit cards fraud detection: Extended abstract. pages 785–786, 10 2020.

[10] Roberto Saia. A discrete wavelet transform approach to fraud detection. 08 2017.

[11] Olena Vynokurova, Dmytro Peleshko, Polina Zhernova, Iryna Perova, and Andrii Kovalenko. *Solving Fraud Detection Tasks Based on Wavelet-Neuro Autoencoder*, pages 535–546. 01 2021.

[12] Badr Omair and Ahmad Alturki. Taxonomy of fraud detection metrics for business processes. *IEEE Access*, 8:71364–71377, 2020.

[13] Badr Omair and Ahmad Alturki. Multi-dimensional fraud detection metrics in business processes and their application. *International Journal of Advanced Computer Science and Applications*, 11, 2020.

[14] Thanasak Ruankaew. The fraud factors. *International Journal of Management and Administrative Sciences (IJMAS)*, 2:01–05, 2013.

[15] Noorhayati Mansor and Rabiu Abdullahi. Fraud triangle theory and fraud diamond theory. understanding the convergent and divergent for future research. *International Journal of Academic Research in Accounting, Finance and Management Science*, 1:38–45, 2015.

[16] Debra D Burke and Kenneth J Sanney. Applying the fraud triangle to higher education: Ethical implications. *J. Legal Stud. Educ.*, 35:5, 2018.

[17] Naqiah Awang, Nur Syafiqah Hussin, Fatin Adilah Razali, Shafinaz Lyana, and Abu Talib. Fraud triangle theory: Calling for new factors. *Editorial Board*, page 1.

[18] David T Wolfe and Dana R Hermanson. The fraud diamond: Considering the four elements of fraud. 2004.

[19] Thanasak Ruankaew. Beyond the fraud diamond. *International Journal of Business Management and Economic Research (IJBMER)*, 7(1):474–476, 2016.

[20] N Christian, YZ Basri, and W Arafah. Analysis of fraud triangle, fraud diamond and fraud pentagon theory to detecting corporate fraud in indonesia. *The International Journal of Business Management and Technology*, 3(4):73–78, 2019.

[21] Yannis Manolopoulos, Charalambos Spathis, and Efstathios Kirkos. Data mining techniques for the detection of fraudulent financial statements. 2007.

[22] Meenatkshi R and Sivaranjani. Fraud detection in financial statement using data mining technique and performance analysis. 9:407–413, 01 2016.

[23] Khaled Gubran Al-Hashedi and Pritheega Magalingam. Financial fraud detection applying data mining techniques: a comprehensive review from 2009 to 2019. *Computer Science Review*, 40:100402, 2021.

[24] Wenkai Deng, Zimin Huang, Jiachen Zhang, and Junyan Xu. A data mining based system for transaction fraud detection. In *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, pages 542–545. IEEE, 2021.

[25] Clifton Phua, Vincent Lee, Kate Smith, and Ross Gayler. A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*, 2010.

[26] Xun Zhou, Sicong Cheng, Meng Zhu, Chengkun Guo, Sida Zhou, Peng Xu, Zhenghua Xue, and Weishi Zhang. A state of the art survey of data mining-based fraud detection and credit scoring. In *MATEC Web of Conferences*, volume 189, page 03002. EDP Sciences, 2018.

[27] Sonika Gupta and Sushil Kumar Mehta. Data mining-based financial statement fraud detection: Systematic literature review and meta-analysis to estimate data sample mapping of fraudulent companies against non-fraudulent companies. *Global Business Review*, 2021.

[28] Eric WT Ngai, Yong Hu, Yiu Hing Wong, Yijun Chen, and Xin Sun. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision support systems*, 50(3):559–569, 2011.

[29] Dianmin Yue, Xiaodan Wu, Yunfeng Wang, Yue Li, and Chao-Hsien Chu. A review of data mining-based financial fraud detection research. In *2007 International Conference on Wireless Communications, Networking and Mobile Computing*, pages 5519–5522. Ieee, 2007.

[30] M Sasirekha, I Sumaiya Thaseen, and J Saira Banu. An integrated intrusion detection system for credit card fraud detection. In *Advances in Computing and Information Technology*, pages 55–60. Springer, 2012.

[31] Tore Dyba, Barbara A Kitchenham, and Magne Jorgensen. Evidence-based software engineering for practitioners. *IEEE software*, 22(1):58–65, 2005.

[32] Mark Staples and Mahmood Niazi. Experiences using systematic review guidelines. *Journal of Systems and Software*, 80(9):1425–1437, 2007.

[33] Barbara Kitchenham and Stuart Charters. Guidelines for performing systematic literature reviews in software engineering. 2007.

[34] Patricia Cronin, Frances Ryan, and Michael Coughlan. Undertaking a literature review: a step-by-step approach. *British journal of nursing*, 17(1):38–43, 2008.

[35] He Zhang, Muhammad Ali Babar, and Paolo Tell. Identifying relevant studies in software engineering. *Information and Software Technology*, 53(6):625–637, 2011.

[36] Babak Darvish Rouhani, Mohd Naz'ri Mahrin, Fatemeh Nikpay, Rodina Binti Ahmad, and Pourya Nikfard. A systematic literature review on enterprise architecture implementation methodologies. *information and Software Technology*, 62:1–20, 2015.

[37] Yuanbang Li, Rong Peng, and Bangchao Wang. Challenges in context-aware requirements modeling: A systematic literature review. In *Asia Pacific Requirements Engeneering Conference*, pages 140–155. Springer, 2017.

[38] Stefan Hoyer, Halyna Zakhariya, Thorben Sandner, and Michael H Breitner. Fraud prediction and the human factor: An approach to include human behavior in an automated fraud audit. In *2012 45th Hawaii International Conference on System Sciences*, pages 2382–2391. IEEE, 2012.

[39] Marco Sánchez, Jenny Torres, Patricio Zambrano, and Pamela Flores. Fraudfind: Financial fraud detection by analyzing human behavior. In *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 281–286. IEEE, 2018.

[40] Namrata Sandhu. Behavioural red flags of fraud—a qualitative assessment. *Journal of Human Values*, 22(3):221–237, 2016.

[41] Jonas Mackevičius and Lukas Giriūnas. Transformational research of the fraud triangle. *Ekonomika*, 92(4):150–163, 2013.

[42] Z Zulaikha, P Hadiprajitno, A Rohman, and R Handayani. Effect of attitudes, subjective norms and behavioral controls on the intention and corrupt behavior in public procurement: Fraud triangle and the planned behavior in management accounting. *Accounting*, 7(2):331–338, 2021.

[43] Normah Binti Omar and Hesri Faizal Mohamad Din. Fraud diamond risk indicator: An assessment of its importance and usage. In *2010 International Conference on Science and Social Research (CSSR 2010)*, pages 607–612. IEEE, 2010.

[44] T Sravanthi, M Sruthi, S Tharun Reddy, T Chandra Prakash, and Ch Vinay Kumar Reddy. Fiscal scam illuminating through analyzing human behaviour. In *IOP Conference Series: Materials Science and Engineering*, volume 981, page 022057. IOP Publishing, 2020.

[45] Shiguo Wang. A comprehensive survey of data mining-based accounting-fraud detection research. In *2010 International Conference on Intelligent Computation Technology and Automation*, volume 1, pages 50–53. IEEE, 2010.

[46] Jianrong Yao, Jie Zhang, and Lu Wang. A financial statement fraud detection model based on hybrid data mining methods. In *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pages 57–61. IEEE, 2018.

[47] R Jayabrabu, V Saravanan, and J Jebamalar Tamilselvi. A framework for fraud detection system in automated data mining using intelligent agent for better decision making process. In *2014 International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE)*, pages 1–8. IEEE, 2014.

[48] Mohiuddin Ahmed and Abdun Naser Mahmood. A novel approach for outlier detection and clustering improvement. In *2013 IEEE 8th Conference on Industrial Electronics and Applications (iciea)*, pages 577–582. IEEE, 2013.

[49] V Kumar and BK Sriganga. A review on data mining techniques to detect insider fraud in banks. *International Journal of Advanced Research in Computer Science and Software Engineering*, 4(12):370–380, 2014.

[50] Ashish Vikram, Sivakumar Chennuru, H Raghav Rao, and Shambhu Upadhyaya. A solution architecture for financial institutions to handle illegal activities: a neural networks approach. In *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*, pages 181–190. IEEE, 2004.

[51] Anwesha Mishra. Fraud detection: A study of adaboost classifier and k-means clustering. *Available at SSRN 3789879*, 2021.

[52] Adila Afifah Rizki, Isti Surjandari, and Reggia Aldiana Wayasti. Data mining application to detect financial fraud in indonesia's public companies. In *2017 3rd International Conference on Science in Information Technology (ICSITech)*, pages 206–211. IEEE, 2017.

[53] Yeonkook J Kim, Bok Baik, and Sungzoon Cho. Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning. *Expert Systems with Applications*, 62:32–43, 2016.

[54] Pediredla Ravisankar, Vadlamani Ravi, G Raghava Rao, and Indranil Bose. Detection of financial statement fraud and feature selection using data mining techniques. *Decision support systems*, 50(2):491–500, 2011.

[55] Prasad Seemakurthi, Shuhao Zhang, and Yibing Qi. Detection of fraudulent financial reports with machine learning techniques. In *2015 Systems and information engineering design symposium*, pages 358–361. IEEE, 2015.

[56] Thakur K & Manju G Mohanty L. Enron corpus fraud detection. *International Journal of Recent Technology and Engineering (IJRTE)*, 8, 2019.

[57] Haibing Li and Man-Leung Wong. Financial fraud detection by using grammar-based multi-objective genetic programming with ensemble learning. In *2015 IEEE Congress on Evolutionary Computation (CEC)*, pages 1113–1120. IEEE, 2015.

[58] Murat Cihan Sorkun and Taner Toraman. Fraud detection on financial statements using data mining techniques. *Intelligent Systems and Applications in Engineering*, 5(3):132–134, 2017.

[59] Abdelali El Bouchti, Ahmed Chakroun, Hassan Abbar, and Chafik Okar. Fraud detection in banking using deep reinforcement learning. In *2017 Seventh International Conference on Innovative Computing Technology (INTECH)*, pages 58–63. IEEE, 2017.

[60] Shahla Mardani, Mohammad Kazem Akbari, and Saeid Sharifian. Fraud detection in process aware information systems using mapreduce. In *2014 6th Conference on Information and Knowledge Technology (IKT)*, pages 88–91. IEEE, 2014.

[61] R Mallika. Fraud detection using supervised learning algorithms. *International Journal of Advanced Research in Computer and Communication Engineering*, 6(6), 2017.

[62] Prajal Save, Pranali Tiwarekar, Ketan N Jain, and Neha Mahyavanshi. A novel idea for credit card fraud detection using decision tree. *International Journal of Computer Applications*, 161(13), 2017.

[63] Carolyn Holton. Identifying disgruntled employee systems fraud risk through text mining: A simple solution for a multi-billion dollar problem. *Decision Support Systems*, 46(4):853–864, 2009.

[64] Mieke Jans, Nadine Lybaert, and Koen Vanhoof. Internal fraud risk reduction: Results of a data mining case study. *International Journal of Accounting Information Systems*, 11(1):17–41, 2010.

[65] Yaya Heryadi and Harco Leslie Hendric Spits Warnars. Learning temporal representation of transaction amount for fraudulent transaction recognition using cnn, stacked lstm, and cnn-lstm. In *2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, pages 84–89. IEEE, 2017.

[66] Suresh Yaram. Machine learning algorithms for document clustering and fraud detection. In *2016 International Conference on Data Science and Engineering (ICDSE)*, pages 1–6. IEEE, 2016.

[67] Samuel Ndueso John, C Anele, O Okokpujie Kennedy, Funminiyi Olajide, and Chinyere Grace Kennedy. Realtime fraud detection in the banking sector using data mining techniques/algorithm. In *2016 international conference on computational science and computational intelligence (CSCI)*, pages 1186–1191. IEEE, 2016.

[68] Jarrod West and Maumita Bhattacharya. Some experimental issues in financial fraud detection: An investigation. In *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, pages 1155–1158. IEEE, 2015.

[69] Chi-Chen Lin, An-An Chiu, Shaio Yan Huang, and David C Yen. Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments. *Knowledge-Based Systems*, 89:459–470, 2015.

[70] Solomon Mekonnen, Keshnee Padayachee, and Million Meshesha. A privacy preserving context-aware insider threat prediction and prevention model predicated on the components of the fraud diamond. pages 60–65, 11 2015.

[71] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

[72] ENRON. (This dataset contains 517,431 emails with 3500 folders from 151 users.).

[73] Emily Homer. Testing the fraud triangle: a systematic review. *Journal of Financial Crime*, ahead-of-print, 12 2019.

[74] Badr Omair and Ahmad Alturki. A systematic literature review of fraud detection metrics in business processes. *IEEE Access*, 8:26893–26903, 2020.

[75] Tahereh Pourhabibi, Kok-Leong Ong, Booi Kam, and Yee Ling Boo. Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems*, 133:113303, 04 2020.

[76] Tore Dybå and Torgeir Dingsøyr. Strength of evidence in systematic reviews in software engineering. pages 178–187, 01 2008.

[77] Tore Dybå and Torgeir Dingsøyr. Empirical studies of agile software development: A systematic review. *Information and Software Technology*, 50:833–859, 08 2008.

# 3 PREDICTIVE FRAUD ANALYSIS APPLYING THE FRAUD TRIANGLE THEORY THROUGH DATA MINING TECHNIQUES

Marco Sánchez-Aguayo[1], Luis Urquiza-Aguiar[2], José Estrada-Jiménez[2]

[1]Departamento de Informática y Ciencias de la Computación, Escuela Politécnica Nacional, Ladrón de Guevara E11-253, Quito 170517, Ecuador

[2]Departamento de Electrónica, Telecomunicaciones y Redes de Información, Escuela Politécnica Nacional, Ladrón de Guevara E11-253, Quito 170517, Ecuador

Email: {marco.sanchez01, luis.urquiza, jose.estrada}@epn.edu.ec

## 3.1 ABSTRACT

Fraud and the losses caused by this phenomenon are increasingly common. There is, thus, an essential economic incentive to study this problem, particularly fraud prevention. One barrier complicating the research in this direction is the lack of public datasets that embed fraudulent activities. In addition, although efforts have been made to detect fraud using machine learning, such actions have not considered the component of human behavior when detecting fraud. In this work, we propose a mechanism to detect potential fraud by analyzing human behavior within a dataset. This approach combines a predefined topic model and a supervised classifier to generate an alert from the possible fraud-related text. Potential fraud would be detected based on a model built from such a classifier. As a result of this work, a synthetic fraud-related dataset is made. Four topics associated with the vertices of the fraud triangle theory are unveiled when assessing different topic modeling techniques. After benchmarking topic modeling techniques and supervised and deep learning classifiers, we find that LDA, random forest, and CNN have the best performance in this scenario. The re-

sults of our work suggest that our approach is feasible in practice since several such models obtain an average AUC higher than 0.8. Namely, the fraud triangle theory combined with topic modeling and linear classifiers could provide a promising framework for predictive fraud analysis.

**KEY WORDS:** fraud triangle; human behavior; topic modeling; data mining; text mining; classification methods.



## 3.2 INTRODUCTION

Fraud is a worldwide phenomenon that affects public and private organizations, including various illegal practices that involve intentional deception or misrepresentation. According to the Association of Certified Fraud Examiners (ACFE), fraud includes any intentional or deliberate act of depriving another of property or money by cunning, deception, or other unfair acts [1].

The 2020 PwC Global Economic Crime and Fraud Survey reports that 49 % of respondents said their companies had been victims of fraud or economic crimes. Approximately 45 % of respondents have experienced losses of less than one hundred thousand dollars; 30 % have suffered losses between one hundred thousand and five million dollars; 6 % have suffered losses between five million and fifty million dollars; and 3 %, losses of more than fifty million dollars. This unveils a rising trend in costs caused by fraud. In organizations, 52 % of cases are related to internal fraud and 41 % to external. This gap is due to anyone in accounting, and financial activities are a potential risk factor for fraud [2].

The prevention of fraud could mitigate expenses related to its prosecution as well as the time and effort to detect fraud after its occurrence. When fraud is discovered, the opportunity to locate the perpetrator and recover the losses caused is scarce. Therefore, organizations must identify factors that lead to fraudulent behaviors and understand their causes. Looking at people in a controlled environment, such as their workplace, we can more accurately identify suspicious behaviors since human behavior analysis is critical in early fraud identification [3].

From the psychological point of view, Donald R. Crassey proposed the fraud triangle theory (FTT) to explain the causes and committing of fraud, identifying the elements that lead the perpetrators to commit fraud. In particular, three elements are represented as the vertices

of the triangle. The fraud triangle's vertices are incentives/pressures, opportunities, and attitudes/rationalization [4]. However, evidence of fraudulent activities in which communications related to this phenomenon are observed is incipient due to its critical and reserved nature, except for certain private and government entities with access to this information. In this context, a valid option is to generate synthetic datasets, which, according to many experts, are the key to making machine learning within artificial intelligence faster and more precise in their predictions [5]. This investigation generated a dataset composed of 14,000 records balanced in two classes of fraud and non-fraud (7000 $\times$ 7000). We identified fraud-related patterns using data mining (DM) techniques and extracted relevant information. On the other hand, relying on text mining (TM) techniques, a subfield of data mining that handles textual data, provides structure to unstructured data. It analyzes it to generate new knowledge [6]. In this context, topic modeling is a widely used approach in TM that provides a comprehensive representation of a corpus by inferring latent content variables called topics. These patterns appear as categories or groups related to content in an unstructured text collection. Therefore, a topic analysis technique assigns a probability to a new text, a document belonging to a specific topic [7]. By calculating the probabilities that a document belongs to a topic, the analysis is performed using classification and deep learning methods to identify which technique is more compatible with topic modeling and efficiently identify phrases suspected of fraud.



To the best of our knowledge, research related to data mining for fraud prediction associated with the fraud triangle theory and its technological applicability is limited or incipient. Auditors detect fraud through the use of their experience, but human bias cannot be easily suppressed, and their reasoning tends to be subjective.

### 3.2.1 Contribution

The main contribution of this work is to propose a novel detector of suspicious behaviors related to the occurrence of fraud by analyzing human behavior using FTT leveraged on machine learning (ML) and deep learning (DL). Our detector combines a predefined topic model and a supervised classifier to alert a potential fraud-related text. In a nutshell, a new document is assigned to the topic of the predefined topic model. In the second step, the text within a topic is classified as a potential fraud-related document, using the topic's probability of the first stage.

We generated a balanced synthetic dataset containing phrases related to fraud and phrases unrelated to fraud. More precisely, the suspicious phrases contain words that belong to a vertex of the fraud triangle (pressure, opportunity, and rationalization). On the other hand, non-fraudulent phrases have a general context that includes words unrelated to this problem. To build our novel detector, we have to do the following:

- Evaluate the performance of text mining techniques, such as Latent Dirichlet Allocation (LDA), non-negative matrix factorization (NMF), and latent semantic analysis (LSA) in the fraud-related dataset. The goal is to select the technique that provides an integral representation of the analyzed documents through clusters, i.e., topic, as separated.
- Once we select the appropriate topic analysis technique, we use the documents' probabilities on the assigned topic to determine if a text can be identified as being fraud-related using supervised machine learning models. For this purpose, we conduct experiments on seven classification methods, including logistic regression (LR), random forest (RF), gradient boosting (GB), Gaussian naive Bayes (GNB), decision tree (DT), k-nearest neighbor (kN), and support vector machines (SVM), using the synthetically generated dataset.
- Furthermore, we perform the same experiment using deep learning techniques, such as convolutional neural network (CNN), dense neural network (DNN), and long short-term memory (LSTM), to determine the performance's differences using receiver operating characteristic (ROC) curves based on the area under the curve (AUC) with the traditional ML classification methods. The goal is to show which technique is more compatible with topic modeling to detect suspicious fraud behavior.

The rest of this paper is organized as follows: Section 3.2.2 presents a literature review in the area of fraud detection. Section 3.3 offers definitions of FTT, topic modeling, classification methods, and deep learning. Section 3.4 describes the data preparation and methodology used in this work. Next, Section 3.5 presents the experiment and the results. Finally, Section 3.6 presents the conclusions and future work.

### 3.2.2 Related Work

Few research papers integrate data mining techniques with analyzing human behavior via fraud triangle theory to identify possible fraud cases. The following studies in the literature

contribute to this topic in this context. In [8], the authors proposed a generic architectural model that considers the fraud triangle factors. In addition to traditional fraud audits, the human factor enhances the audit analysis since the transactions examined by an auditor can be differentiated and prioritized better. By distinguishing behaviors (suspicious and non-suspicious), it is possible to discover transactions that are part of a pattern that is not yet known and that would have been left undiscovered if only traditional means were used. Likewise, Carolyn Holton in [9] proposed the design of a detector of disgruntled communications, mainly in email repositories, through data mining techniques associated with the triangle of fraud theory to combat internal security risks. In these lines, Mieke Jans [10] focused on reducing the risk of internal fraud by combining the detection and prevention of fraud. Its analysis uses descriptive data mining techniques to identify whether an observation is fraudulent or not. In this investigation, the authors used the IFR$^2$ methodology [11] to reduce the risk of internal fraud, a framework that uses the fraud triangle theory to assess and minimize fraud opportunities. Vimal Kumar [12] analyzed fraud in the banking sector, classifying the types and definitions of existing fraud mechanisms. He also listed and explained the different data mining techniques used by investigators to study fraud, taking into account the factors that cause it by using the fraud triangle model, relating the pressure, timing, and rationalization with this behavior. The author concluded that prevention is an indispensable requirement in the banking sector, and data mining techniques are essential to reduce fraud cases. Ravisankar [4] used the fraud triangle theory to identify the possible reasons for increased fraudulent activities in companies. The authors used the multilayer feed-forward neural network (MLFF), support vector machines, group method of data handling (GMDH), genetic programming (GP), logistic regression (LR), and probabilistic neural network (PNN) to predict fraud in financial statements on a dataset from 202 Chinese companies. Their results showed that PNN was the technique with the best performance, followed by GP. Aside from the methodology proposed by Jans [11], some other frameworks for fraud detection have been proposed. Panigrahi [13] suggested the integration of an auditor knowledge base and the techniques in audit processes called “knowledge-driven internal fraud detection (KDIFD)” to help auditors in the discovery of internal financial fraud more efficiently by applying data mining techniques. Authors in [14] proposed a system based on an automated framework for fraud detection using intelligent agents, data fusion techniques, and various data mining techniques. Works on revisions of data mining techniques and machine learning applied to fraud detection were also identified, as in the case of [15] in which the authors reviewed research works on the methods of data mining applied to financial fraud detection

(FFD). In [16], the authors classified, compared, and summarized fraud detection methods and techniques based on mining relevant data in published academic and industrial investigations. This work also highlighted the application of data mining in other related fields, such as epidemic detection, insider trading, intruder detection, money laundering, spam detection, and terrorist detection. In the same context, Wang et al. [17] reviewed the literature on data structure algorithms. The authors provided a reference to optimize fraud detection models. They aimed to collaborate with public accountants to select data and data mining technologies suitable for detecting fraud. Dhiya Al-Jumeily [18] compared existing systems for fraud detection and proposed developing a new system that allows the detection of potentially fraudulent applications. With this method, organizations have a good outlook on the authenticity of applicantsídentities and online applications. On the other hand, the problem of the lack of access to financial data for fraud investigation has been addressed by other works using simulation techniques; thus, the privacy concerns of accurate data are avoided. Lopez et al. presented in [19] three case studies related to financial transactions, where a method to generate synthetic data was offered, which can be used as part of the necessary input data for the research, development, and testing of fraud detection techniques. Similarly, ref. [20] proposed a novel way to create synthetic data for fraud investigation by developing a simulation prepared with accurate data. In [21], simulation techniques aimed to recreate the behavior of fictitious clients. All the reviewed works contribute to fraud detection, mainly in the banking sector, proposing reference frameworks, such as IFR$^2$ and even applications related to artificial intelligence. However, the fraud analysis focused on a semantic context trying to identify unusual patterns in a dataset is still incipient. Moreover, the previously mentioned articles did not address the combination of text mining with the fraud triangle theory to categorize texts as potentially fraud-related. In this sense, no studies were identified with evidence of the use of data mining techniques, the application of fraud theories, and the corresponding analysis of human behavior to detect fraud, which means that there is a gap, and this is an appropriate field investigation.



## 3.3 MATERIALS AND METHODS

This section briefly describes the fraud triangle theory, topic modeling strategy, classification methods, and validation methods.

### 3.3.1 Fraud Triangle Theory (FTT)

Fraud is considered a subset of internal threats, such as corruption, misappropriation of assets, and fraudulent statements, among others[22]. ACFE defines fraud as “the use of one’s occupation for personal enrichment through the deliberate misuse or misapplication of the employing organization’s resources or assets” [23]. There are two types of fraud: internal and external. Internal fraud covers a series of irregularities and illegal acts characterized by the scammersíntentional deception, leading to the misappropriating a company’s money and other essential resources. In the case of external fraud, this is commonly done in the financial statements, which are falsely presented in the reports [13]. The fraud triangle theory proposed by Donald R. Cressey comprehensively explains this phenomenon’s occurrence. Cressey, a leading sociology expert, wrote several books on preventing this crime. The reasons for committing it could be summarized in the following three critical elements: perceived pressure, opportunity, and rationalization. This theory determines that all three parts must be consecutively present to suspect a desire to commit fraud. The pressure is what motivates the crime in the first place. For instance, the subject has some economic problems that he cannot solve by legitimate means, so he begins to consider committing an illegal act, such as stealing cash or forging financial statements, to solve his problem [24]. The second element is the perceived opportunity, which defines how the person will commit the wrongful act. The person must see how he can use (abuse) their position of trust to solve their financial problems with a low perception of the risk of being discovered. Finally, the third component relates to the idea that individuals can rationalize dishonest actions. Most people who commit fraud do it for the first time and do not have a history of criminality. They see themselves as normal, honest people who have come up with a series of situations. Consequently, the fraudster will justify his actions in a way that is acceptable [18]. The risk of committing fraud increases when there is a tight connection between pressure, opportunity, and rationalization.



### 3.3.2 Topic Modeling (TM)

TM is commonly applied to extract valuable knowledge when performing text mining. TM allows the identification of hidden semantic structures related to a particular “topic”. TM analyzes collections of documents, representing each as a mix of topics. In turn, a proba-

bility distribution over the words contained in the documents models each topic [25, 26]. If a document is about a specific topic, the words related to that topic will be present more frequently than the others. For example, a chat message about the poor economic situation of a person (potentially related to the pressure component of the fraud triangle) may contain words such as “debts”, “financial problems”, and “late payments”. Three unsupervised machine learning algorithms are commonly used to implement topic modeling: LSA, NMF, and LDA [27]. From the evaluation point of view of TM methods, the key metrics are perplexity [25] and coherence to select an adequate number of topics depending on the problem at hand. The perplexity value is a confusion metric and accounts for the level of “uncertainty” in a model’s prediction result. In contrast, the coherence score indicates the level of semantic similarity between words on a topic [28]. In this sense, coherence provides a more decisive factor in parameter optimization for this work, which is why this metric was chosen to analyze topics [29].



#### 3.3.2.1 Latent Semantic Analysis (LSA)

LSA is a technique that allows us to create a vector representation of texts to create semantic content. Through this “vector” representation, LSA calculates the similarity between texts to choose the most accurately related words. LSA uses singular value decomposition (SVD) to reduce the vector space dimensions. LSA tries to capture the latent semantics in linear space [30]. The idea is to obtain vectors for each document so that we can use them to find similar words and similar documents [31]. LSA collects a large amount of text, divides it into documents, and then creates a matching matrix of terms and documents through SVD.

#### 3.3.2.2 Non-Negative Matrix Factorization (NMF)

Provided a set of $n$ documents, $m$ unique words and $k$ topics, NMF unveils the main hidden themes by decomposing the non-negative matrix of term-documents $D \in \mathbb{R}_{+}^{m \times n}$ in the product of two other matrices; one matrix $U \in \mathbb{R}_{+}^{m \times k}$ that represents the relationships between words and themes and matrix $V \in \mathbb{R}_{+}^{k \times n}$ encloses the topic–document information in the latent topic space (i.e., $D \approx UV$) [32]. NMF is a form of dimension reduction because the number of topics *k* is typically many orders of magnitude smaller than the number of words *m* and several documents *n* under consideration. Matrices *U* and *V* constitute the principal

result of NMF, and the distribution of words and documents about the topics is the primary focus of interpretation [33].



#### 3.3.2.3 Latent Dirichlet Allocation (LDA)

LDA is an unsupervised probabilistic generative model that allows finding the semantic structure of a corpus. LDA is based on the hierarchical Bayesian analysis of texts [34]. An LDA model considers several themes in a corpus and a document as a bag of words generated from these themes. In LDA, each document is modeled as a random mix of latent topics. In turn, each topic is characterized as a probability distribution over words; that is, each vocabulary word has a certain probability, where words with high probability are more associated with that topic than words with low probability [35].

A word is defined as a basic unit of discrete information, which will be part of a vocabulary that we can denote as $\{w_1, w_2, \ldots, w_V\}$. A document is a sequence of words represented by $\{w_1, w_2, \ldots, w_N\}$, where $N$ denotes the number of words present in the document. A corpus $D = \{d_1, d_2, \ldots, d_M\}$ is a collection of documents that includes the texts on which the topic analysis is to be carried out, and $M$ is the number of documents in the corpus. The $K$ topics present in the corpus are represented by the vector $\beta$. The $k$ topic (i.e., $\beta_k$) can be considered a distribution over the vocabulary. To the presence of the $k$th topic in a particular document $d$, we call it $\theta_{d:k}$. The assignments of a word $n$ of a document $d$ in a specific topic are denoted as $z_{d,n}$. Finally, the words observed in a document $d$ are denoted as $w_d$, and particularly the $n$-th word of the document is denoted as $w_{d,n}$. More formally, Ref. [36] defined these dependencies in the generative process for LDA, which depicts the joint distribution of hidden and observable variables in the model, as can be seen, in Equation (3.1).

$$
\begin{aligned}
& p\left(\beta_{1:K}, \theta_{1:D}, w_{1:D}\right) \\
& \qquad = \prod_{k=1}^{K} p\left(\beta_k\right) \prod_{d=1}^{D} p\left(\theta_d\right) \\
& \qquad \left( \prod_{n=1}^{N} p\left(z_{d,n} \mid \theta_d\right) p\left(w_{d,n} \mid \beta_{1:K, z_{d,n}}\right) \right)
\end{aligned}
\tag{3.1}
$$

Figure 3.1 illustrates a probabilistic graphical model (PGM), where the conditional dependencies of the different variables involved in the generative process of the LDA algorithm are observed. The white nodes represent latent variables, such as the prevalence of each

topic in a document, the assignment of each word in the document to a topic, and the topics themselves. The shaded node represents the unhidden and observable variable.

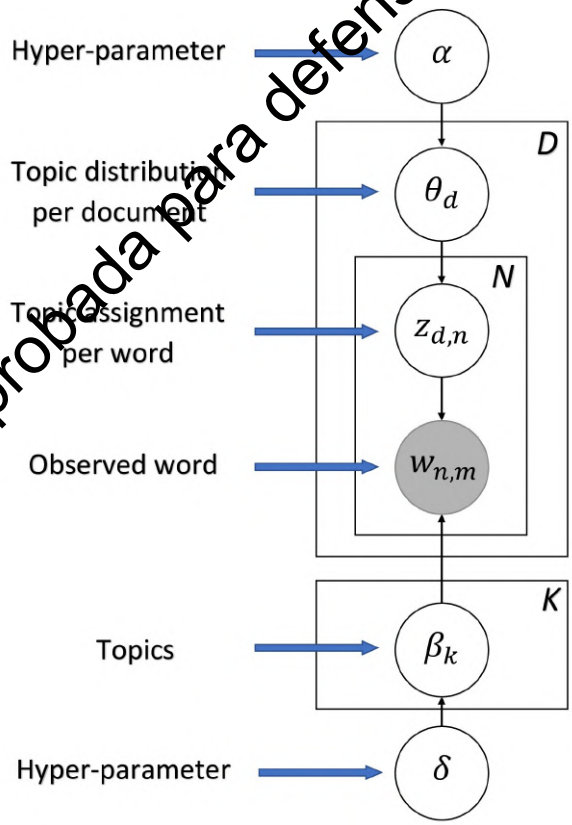




**Figure 3.1:** Representation of latent Dirichlet allocation LDA. Hidden nodes are not shaded and represent the proportions of topics, assignments, and topics.

The computational problem is to compute the conditional (posterior) distribution of the topic structure, according to Equation (3.2).

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(_{1:D})} \tag{3.2}$$

The numerator is the joint distribution of all the random variables, which can be easily computed for any hidden variable settings. The denominator is the marginal probability of the observations: the probability of seeing the observed corpus under any topic model.

### 3.3.3 Classification Methods

Supervised machine learning (ML) classifiers have several applications, including predictive data mining. These algorithms carry out the assignment of objects into labeled classes or

categories of information. Classification is a supervised machine learning approach that consists of labeling data items as belonging to a particular class from a model built from a selected dataset. In other words, a training dataset is used to derive a model, which is then used in the new datasets to classify unseen test data [37]. In this context, the classifier observes a set of training samples and, following in them, can make predictions about the categorization of some other new samples presented.

Each classifier has an associated precision that will differ according to the type of data used. There are several evaluation metrics to compare the classification methods, and each of them could be useful depending on the kind of problem associated [38]. A receiver operating characteristic (ROC) chart is a technique for visualizing, organizing, and selecting classifiers based on their performance. The area under this curve (AUC) is one of the most critical evaluation metrics that can be applied to choose the best method of classification [39]. In addition, AUC is among the most commonly used performance metrics in the literature related to fraud detection. Thus, for comparability with other works, we decided to apply this metric to assess the models obtained in this work.

This work compares six well-established classification algorithms to detect fraud-related text within a topic using the AUC criterion.



#### 3.3.3.1 Logistic Regression (LR)

LR, also known as a logistic or logit method, analyzes the relationship between multiple independent variables and a categorical dependent variable, estimating the probability of occurrence of an event by fitting the data from a logistic curve.

#### 3.3.3.2 k-Nearest Neighbor (kN)

The kN algorithm was developed from the need to perform discriminant analyses with unknown parametric estimates of probability densities [40]. kN can classify unlabeled observations by assigning them to the class of the most similar labeled examples [41]. There are two essential factors related to this classifier. One is the method for calculating the distance between a sample and others belonging to the most frequent class among the closest training examples. In most cases, the kN implementation uses the Euclidean distance. The other factor is to decide how many neighbors (i.e., $k$) to choose in this algorithm.

#### 3.3.3.3 Decision Tree (DT)

The DT algorithm solves classification and regression problems in the form of trees. DT can be updated incrementally by dividing the dataset into smaller datasets (numerical and categorical), where the results are represented in the leaf nodes [42]. Decision trees are generally represented as a hierarchical structure that allows a more accessible interpretation than other methods; each internal node checks an attribute, while each branch corresponds to the attribute's value or range of values [43].

#### 3.3.3.4 Random Forest (RF)

RF is a classification method based on several decision trees, which is used to classify a new instance by majority vote. Each node in the decision tree uses a subset of attributes selected randomly from the entire original set of characteristics [44]. The correlation between trees decreases by randomly selecting the features that improve the predictability, and higher efficiency is obtained as a result [45].

#### 3.3.3.5 Gaussian Naïve Bayes (GNB)

The GNB classifier applies Bayes'theorem, assuming that all attributes are independent. Its main advantage is that it requires a small measure of training data vital for the characterization and necessary for classification [46]. The GNB classification is a case of the naive Bayes method, assuming that there is a Gaussian distribution on the attribute values, given the class label.

#### 3.3.3.6 Gradient Boosting Decision Tree (GBDT)

The GBDT is based on the decision tree model. It builds the model through gradient augmentation, aiming to boost the combination of several weak and simple classifiers in a given set. This algorithm trains a new tree model that reduces the error of the whole set. To ensure that the loss function decreases continually in each iteration, the new tree model is built using the loss function's negative gradient [47]. Compared with linear regression models, GBDT can handle different types of variables (continuous, categorical, etc.) and requires

little data preparation time [48].

#### 3.3.3.7 Support Vector Machines (SVM)

Vapnik introduced SVM as a kernel-based machine learning model for classification and regression tasks. SVM classifier aims to find a linear hyperplane (decision boundary) that separates the data to maximize its margin. For example, look at a problem of separating two classes in two dimensions [49, 50]. SVM is a high-precision binary data classification technique that has been widely used in various fields. Let $v = \{v_1, v_2, \ldots, v_m\}$, an $m$-dimensional input feature. We assume that each $v_i \in v$ is normalized to an interval $[0, n]$ using normalization techniques. Let $p = \{-1, +1\}$ be two different predictions that is, negative and positive. An SVM classifier is a separating hyperplane with a maximum margin in the $m$-dimensional feature space, which divides the $m$-dimensional feature space into two subspaces, i.e., a subspace for positive prediction and the other for negative prediction [51].



### 3.3.4 Neural Networks

A neural network (NN), also known as an artificial neural network (ANN), allows non-linearity between the characteristic variables and the output signals [52]. A simple NN generally consists of an input layer, a hidden layer (s), and an output layer. The number of hidden and output layers is the neural network depth. The term deep learning refers to NN with considerable depth [53]. In this investigation, the input layer receives the information of the document probabilities. These belong to a specific topic; the output layer predicts the result, in this case, whether or not the sample is associated with fraud cases.

#### 3.3.4.1 Deep Learning (DL)

DL is a subfield of machine learning in artificial intelligence based on algorithms that try to model high-level abstractions in data through the use of multiple layers of processing with complex structures or composed of multiple non-linear transformations [54]. TensorFlow, Keras, and PyTorch are the most used libraries for DL. For this work, we will use Keras, a high-level framework written in Python that provides a second-level abstraction, that is, instead of directly using the first-level frameworks (Theano, Torch, PyTorch, and Tensorflow).

We can use a new framework over an existing one and thus further simplify the development of the deep learning model.



### *Dense Neural Networks*

Dense neural networks (DNN) are also known as feed-forward networks because they avoid cycle formation. Determining the adequate number of neurons in hidden layers is a complicated issue (done by trial and error) since many neurons can result in overfitting problems. In contrast, a small number cannot learn from the data [55].

At the output of each layer, we have an activation function. In the next layer, the value of one of the neurons corresponds to the image of the values of the previous neurons, representing the non-linearity in a neural network. The output is composed of the selected activation functions, commonly non-linear ones, such as sigmoid, hyperbolic tangent, and ReLU, among others [56].

### *Convolutional Neural Networks*

A convolutional neural network (CNN) is a deep learning algorithm that allows processing data with local patterns, which is very efficient for image classification [57]. It comprises an input layer, convolutional layers, and fully connected layers on top. Additionally, it uses tied weights, grouping layers, and an output layer. This architecture allows CNNs to take advantage of the 2D structure of the input data [54].

### *Long Short-Term Memory*

Long short-term memory (LSTM) is an improvement of the recurrent neural network (RNN), which has the problem of the gradient's disappearance or explosion. LSTM memory blocks are used instead of conventional RNN units to solve this problem. RNN sometimes fails to capture long-term dependency in a sequence. Therefore, short-term memory was invented to solve this problem by recursively applying a transition function to the input's hidden state, allowing it to remember and connect the previous information to the current one [58].

LSTM retains a cell state $C_t$ in the time interval $t$ that allows it to learn the formed, stable

sequential correlations. LSTM controls information flow through the entry gate, forgetting gate, and exit gate [59].



## 3.4 METHODOLOGY FOR PREDICTING FRAUD BASED ON THE FRAUD TRIANGLE COMPONENTS

Our objective is to build predictive models to enable early fraud detection. Thus, our strategy consists of identifying hidden patterns that might be related to one of the fraud triangle vertices from the fraud triangle theory. For this, we construct a model to predict whether a specific phrase belongs to one of these triangle categories. In this line, this strategy is a novel approach to fraud detection as long as it considers a new semantic view of this problem.

To detect suspicious patterns related to the vertices of the fraud triangle, we first perform topic modeling (unsupervised learning) over an unstructured text data set [60]. In particular, we select the best model obtained from LSA, NMF, and LDA.

Then, based on the coherence value, we determine the appropriate number of topics we can align with the fraud triangle theory; this involves obtaining the probabilities of documents belonging to a specific topic.

Based on such topics (or labels) and machine learning techniques, we categorize a sentence as potentially fraudulent if there is suspicion of it belonging to one of the vertices of the fraud triangle. This process is illustrated in the first flow chart of Figure 3.2.

Once the probability that a document belongs to a specific topic is calculated through topic modeling, a balanced dataset is obtained with records labeled as fraud and non-fraud. This dataset allows training learning models to predict potentially fraudulent behavior. This is illustrated in the second flow chart of Figure 3.2.

The performance of supervised learning methods applied over this data set is benchmarked to identify the best-performing one. Finally, the results obtained are analyzed to determine which technique is most compatible with topic analysis for fraud identification. More details on the process followed to implement this strategy are provided below.

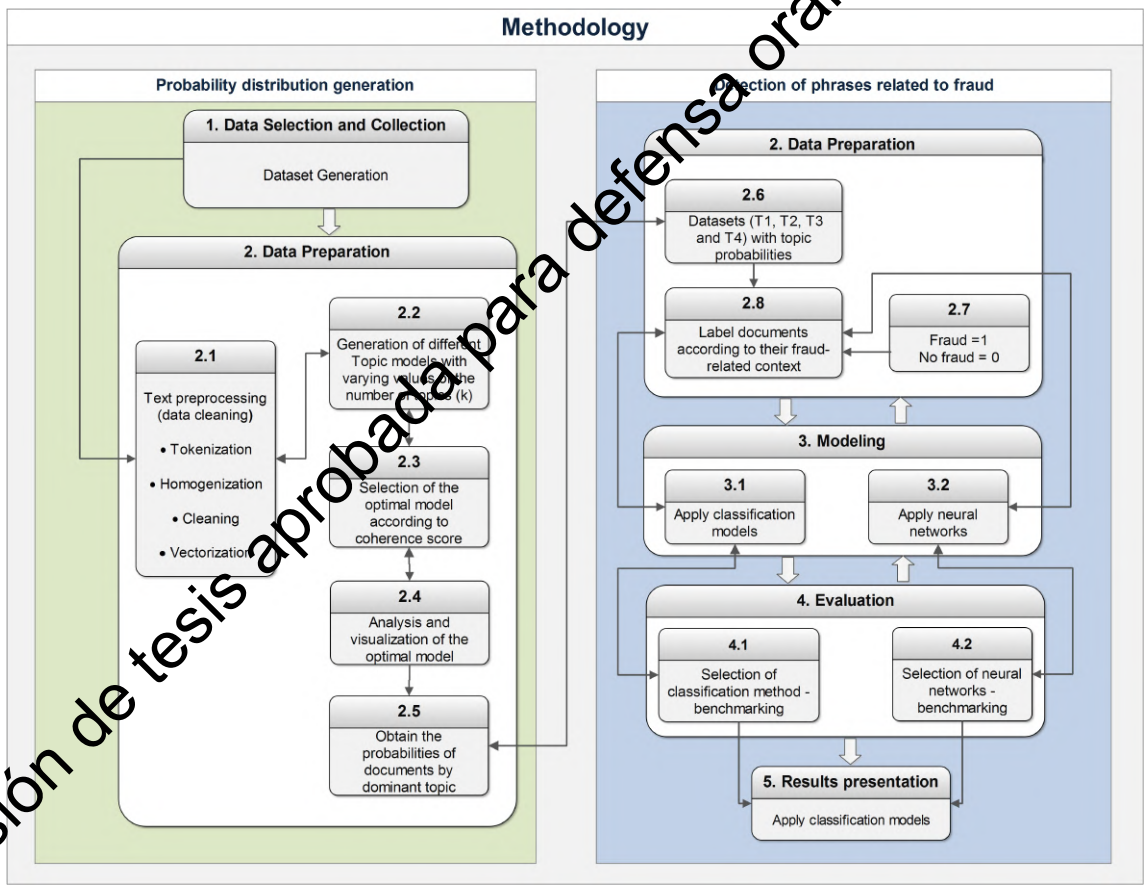




**Figure 3.2:** Methodology used to determine the existence of fraud.

### 3.4.1 DataSet Generation

Datasets involving fraud-related behavior are scarce due to several reasons, e.g., due to the confidentiality policies of institutions or due to the sensibility of the personal data included. Given the restricted access to this information, it is common to use synthetically generated datasets [61, 62]. Our dataset was created from a dictionary of fraud-related keywords that were purchased from the company [63]. These keywords are tagged into different categories, including pressure, rationalization, and opportunity, the three components of the fraud triangle theory. Starting from several of these keywords related to the fraud triangle theory and using different online tools to generate sentences, as in [64, 65, 66], the corresponding sentences, including the selected keywords, were obtained. These tools allow the generation of sentences based on a specific word with a well-defined grammatical and semantic structure. Finally, they use a web scraping tool, “Firefox Addon,” which allows us to save the generated results and export them in CSV format for processing. The process followed to generate the dataset is shown in Figure 3.3. Additionally, following the same procedure,

several documents not related to fraud were generated in the same proportion as those related to fraud, with the only difference being that, for this case, keywords not related to this phenomenon were chosen, thus obtaining a balanced dataset.

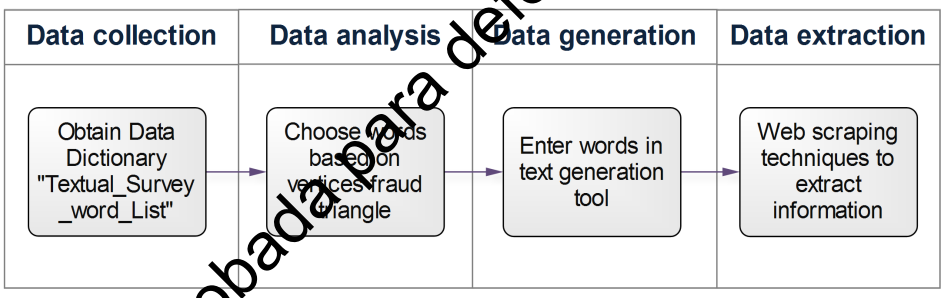


**Figure 3.3:** Flow diagram used for the generation of a synthetic dataset.

The next step is to analyze the data and identify their characteristics to classify their main parameters, which must be retained for the early detection of suspicious behavior related to fraud.



### 3.4.2 Data Preprocessing

The raw representation of the dataset needs to be changed to be more suitable for topic modeling. With this aim, we use NLTK (Natural Language Toolkit), a Python library that implements some of these preprocessing phases: sequential tokenization, homogenization, cleaning, and vectorization [67]. Next, we describe them.

#### 3.4.2.1 Tokenization

Tokenization is key for text processing. It consists of changing the representation of the dataset so that it can be more easily processed. With this aim, tokenization involves dividing a document into its words (tokens). This was implemented using Python through the `word_tokenize` function [68] from the nltk.tokenize package.

#### 3.4.2.2 Homogenization

Homogenization entails adapting the dataset by eliminating certain parts of words and sentences that do not contribute to the semantic analysis of the text. Some of these activities

are described below.

a . Change all tokens to lowercase. This is implemented by the Python `lower` function.

b . Remove non-alphanumeric items. To identify non-alphanumeric characters, we use the Python `isalnum` function.

c . Obtain the word lexeme (lemmatization). Lemmatization turns words into their lemma/lexeme form (for example, “runs”, “running”, and “ran” are all forms of the word run, and therefore “run” is the lemma of all these words). When obtaining lexemes, word sets are uniquely represented. In this way, the semantic meaning of the words is associated with the same lexeme. For this, we use the `lemmatize` [69] function of WordNetLemmatizer from NLTK.



#### 3.4.2.3 Cleaning

It is essential to mention that there will be sets of words that do not add semantic value to documents. The cleaning process is based on eliminating less relevant words, that is, those that provide less information. For example, articles, prepositions, or conjunctions are words of little relevance. The `stopwords list` function [70] provided by NLTK is used to identify these words and to remove them.

#### 3.4.2.4 Vectorization

Vectorization entails obtaining a numerical representation of the words or phrases of a dataset. Vectorization aims to extract more useful information when processing natural language text, e.g., through LDA.

LDA topic modeling requires vectorized documents. To implement vectorization, we use the `gensim` library. The dictionary function belonging to this library allows building a dictionary containing all the tokens that appear in the corpus and assigning them an identifier. We use both this dictionary and the function `doc2bowse` [71] that converts documents to a word bag representation. The corpus is constructed in the format necessary to carry out topic modeling through algorithms that implement LDA.

### 3.4.3 Quantitative Evaluation of Topic Modeling Algorithms

The most relevant and used topic modeling methods are LSA, NMF, and LDA. In related research, it was observed that the effectiveness of these algorithms differs in terms of the amount and type of data to be processed. In most cases, and particularly for large datasets, LDA proved to be more efficient than other methods when identifying coherent topics [72, 73]. In more specific cases, NMF outperformed the others [74]. In general, NMF and LDA are similar, but LDA seems to be more consistent [75]. Although the use of LDA has become popular when handling big unstructured data, selecting the best option for topic modeling might depend on the particular data being processed. Consequently, benchmarking the efficiency of these algorithms in this context is required. First, we identify the appropriate number of topics based on the resulting coherence value of each model. This approach enables us to analyze the performance of the topic mentioned above modeling algorithms and, in particular, to identify the one that more concisely and coherently learns such topics. Having identified the $k$ parameter (number of topics) for models obtained from LDA, NMF, and LSA, we select the algorithm offering the highest coherence value, which identifies the end of the rapid growth of coherence between topics, thus offering meaningful and interpretable topics. From this analysis, a quantitative evaluation of the topic modeling techniques is carried out, which consists of measuring the coherence of the topic $C\_v$ over a model's topic and topic–article assignment output, which will indicate an approximate measure of the quality of that result.



### 3.4.4 Selection of the Topic Modeling Algorithm

After obtaining topic models based on LDA, NMF, and LSI from the dataset, we evaluate the consistency of the sets of words generated by each and determine the efficiency of classifying them into a specific topic. Each topic groups the most representative words for a given subject. We compare the sets of words obtained by the models and their distribution, prevalence, and structure. This analysis enables us to find the method that more accurately identifies the dataset's topics. Once the model with the best performance is identified, it is checked whether its value $k = 9$ is the most appropriate. If this is the case and non-overlapping topics are found, that is, a coherent structure of particular topics, this would be an appropriate value of k. Otherwise, we will manually identify another value of $k$ that meets

a suitable distribution of topics based on visual inspection of the topics found.

Afterward, we obtain the modeling of the topics corresponding to this $k$ value. Here, we analyze the distribution of words corresponding to each topic and identify the words related to fraud that are more representative or dominant; the objective is to find a relation between the context of each topic and the vertices of the fraud triangle. Then, from the LDA model, we obtain the probabilities that the documents in the data set belong to a specific topic. Each of these probabilities may represent a metric to categorize a document as being potentially related to fraud. Nevertheless, such probabilities themselves also serve as an interesting new representation of the dataset. From this representation, we extract smaller datasets, each of which groups documents associated with a (dominant) topic, i.e., a topic to which the documents belong with the highest probability.



### 3.4.5 Methodology of Evaluation

Because the dataset was synthetically generated, it is possible to identify a priori fraud-related phrases, label them accordingly, and then show how accurate a classification method is when predicting fraud activities. It is necessary to identify which model best fits the analysis of topics in the context of the dataset, its size, and its characteristics. When analyzing the performance of traditional machine learning and deep learning models, traditional classifiers can generally learn better than deep learning classifiers if the dataset is small. On the other hand, deep learning models might obtain a performance boost when working over larger datasets. We evaluate both approaches since the intrinsic characteristics of a dataset could affect their performance.

Once behavior patterns related to fraud are identified through the topic analysis and probability distributions generated, we have a dataset that can be analyzed using classification methods and neural networks. We aim to evaluate how accurate the prediction of these models turns out to be. The most common classification and deep learning methods are used to identify which alternatives have better performance.

The analysis of both techniques is carried out using the ROC curve graph; this allows us to visualize, organize, and select classifiers based on their performance using the AUC parameter that depicts the quality of classification methods.

## 3.5 RESULTS AND DISCUSSION

This section presents the results obtained from testing our fraud detection mechanism in a case study. From our view, such results show its effectiveness. Details on data collection and processing are provided, followed by experiments on supervised and unsupervised model learning and the analysis of such results. Finally, the practical implications of the method and the findings are discussed.



### 3.5.1 Probability Distribution Generation

In this first scenario, we present the results from analyzing our synthetic dataset to find patterns related to the fraud triangle theory, which is the proxy we use to detect potential fraud-related behaviors.

#### 3.5.1.1 Optimal Number of Topics

When topic modeling is used, it is essential to determine the number of topics ($k$) that best capture the trends in potentially fraudulent messages. We constructed several models based on LSA, NMF, and LDA with different values of $k$, and those with the highest coherence score were selected. Choosing several $k$ topics associated with the maximum resulting coherence generally offers the most appropriate topics.

To obtain the coherence value of different models, `scikit-learn` and `gensim` Python libraries were used. Gensim does not have an implementation of NMF, so it was used only to implement LDA and LSA. `scikit-learn` offers a solution for NMF, allowing it to obtain the required coherence value.

The coherence validation for the different numbers of topics is shown in Figure 3.4. In the three models (LSA, NMF, and LDA), we can observe that the coherence value increases as the number of topics increases, demonstrating that the patterns in data are better captured with a higher number of topics. For the three models, the coherence value gradually increases to a certain $k$. For LSA, we obtained the highest coherence value when $k = 4$. For NMF, $k = 8$ resulted in a coherence value of $0.9143$. LDA obtained the highest coherence value ($0.6164$) for $k = 9$. For higher values of $k$, for all three cases, coherence fluctuates indetermi-

nately. This result implies that establishing a higher number of topics does not necessarily imply better performance. Instead, the time necessary for their calculation increases.

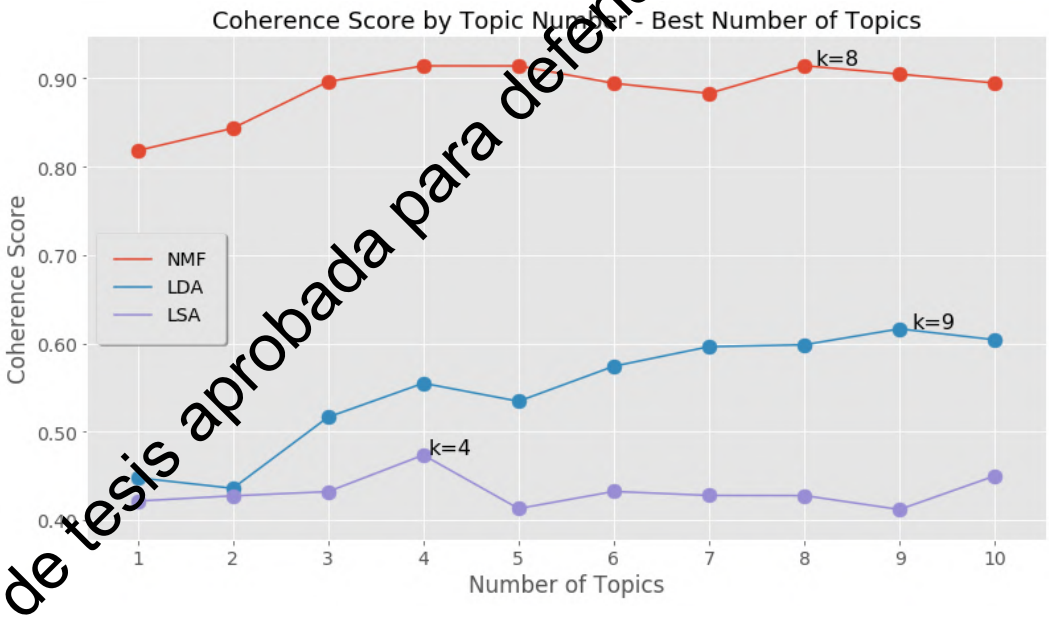


**Figure 3.4:** Comparing the techniques (LSA, NMF, and LDA)—highest coherence score.



NMF showed the highest coherence score, followed by LDA and LSA, respectively. NMF classified large numbers of phrases on a specific topic. On the other hand, LDA was able to better distribute phrases along with all $9$ topics, according to Table 3.1.

**Table 3.1:** Highest values of coherence obtained from the three models.

| | **Models** | | |
|---|---|---|---|
| | **LSA** | **NMF** | **LDA** |
| Coherence Values | 0.4735 | 0.9143 | 0.6164 |

To analyze the behavior in the distribution of the topics, based on the values of $k$ we tested, Tables 3.2–3.4 show the 10 most relevant keywords of the analyzed models associated with their related topic. Four topics were discovered using LSA, and some of the words grouped in topics overlap; this is because the dimension of latent themes depends on the range of the corresponding matrix (see Section 3.3), and this limit is exceeded. Additionally, LSA cannot capture the different meanings of words, offering less precision when distributing words in each topic. Table 3.2 depicts how words are clustered into topics (context) when using LSA and illustrates how some words are repeated for some topics, which is evidence of the problem of capturing the meaning of words. We use words in color to visually represent this

phenomenon.

**Table 3.2:** Collection of topics and the top 10 keywords of the corresponding topic represented by the LSA model.

| LSA | | | |
|---|---|---|---|
| **T1** | **T2** | **T3** | **T4** |
| problem | debt | be | job |
| economic | public | scare | lose |
| debt | problem | job | be |
| social | economic | lose | scare |
| political | country | go | ill |
| face | private | know | would |
| solve | service | get | scared |
| country | include | care | want |
| serious | reduction | think | work |
| issue | stock | people | earning |
| people | total | deserve | get |



In the case of NMF, the overlapping of words along topics is also evident, as depicted in Table 3.3. Since more topics are involved with NMF, the repetition of words would have a less negative impact than with LSA. Note that the repetition of words may also be due to a too-high value of $k$.

Finally, as shown in Table 3.4, the LDA model best groups words in topics since none of such words are repeated; this might entail a more consistent distribution of words along topics.

Since LDA behaves better on topic modeling in this particular context, we next evaluate this algorithm to detect potential fraud activities.

#### 3.5.1.2 Application of LDA Model

The number $k$ of topics is an input parameter to obtain an LDA topic model. Determining the adequate value of $k$ is critical for the model's performance. For our particular scenario (fraud detection using the fraud triangle theory), intuitively, the ideal number of topics embedded in the dataset is 3, corresponding to the vertices of the fraud triangle (pressure, opportunity, and reasoning). However, from the coherence analysis described previously, $9$ is the excellent value of $k$.

Such overlapping could also be analyzed through an intertopic distance map, e.g., that provided by the `pyLDAvis` Python library. `pyLDAvis` depicts an interactive, visual representation of an LDA model through bubbles that represent the topics in a semantic topic space. Then, the closer the bubbles are to each other, the more semantic similarity they share. This map facilitates the understanding of the topic-term relationships in an adjusted LDA model and

**Table 3.3:** Collection of topics and the top 10 keywords of the corresponding topic represented by the NMF model.

| NMF | | | | | | | |
|---|---|---|---|---|---|---|---|
| **T1** | **T2** | **T3** | **T4** | **T5** | **T6** | **T7** | **T8** |
| debt | economic | tom | system | scared | review | job | easily |
| public | problem | mary | failure | people | period | lose | accessible |
| external | problem | big | error | know | currently | get | hotel |
| countries | social | think | file | got | keep | want | public |
| sustainability | political | want | data | really | kept | temporary | transport |
| private | issue | know | case | something | matter | steal | information |
| restructuring | serious | going | power | think | committee | work | car |
| total | countries | told | due | away | earnings | deserve | bus |
| reduction | people | help | event | look | countries | going | foot |
| management | country | thought | computer | get | board | need | city |

**Table 3.4:** Collection of topics and the top 10 keywords of the corresponding topic represented by the LDA model.

| LDA | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **T1** | **T2** | **T3** | **T4** | **T5** | **T6** | **T7** | **T8** | **T9** |
| steal | review | poor | want | people | big | make | economic | problem |
| later | think | child | deadline | know | use | care | weakness | debt |
| support | time | need | failure | evacuation | exploitation | job | ill | fair |
| say | fix | inadequate | year | deserve | right | work | life | abuse |
| just | help | insufficient | temporary | unethical | labor | compensation | leave | easily |
| tell | come | country | day | issue | family | lose | feel | accessible |
| woman | look | supervision | man | cause | friend | good | face | case |
| live | scare | really | old | situation | different | earning | thing | car |
| currently | like | money | ask | away | girl | way | great | information |
| period | world | school | change | abuse | hope | new | social | food |



offers additional information about other perspectives on the applied model [76].

We tested the intertopic distance map in Figure 3.5 for different values of $k$, and we found an evident overlapping of topics for $k = 9$. In contrast, for $k = 4$, this representation depicted in Figure 3.5b showed topics adequately separated from each other. These four topics would be in line with the categories associated with the three components of the fraud triangle plus a fourth topic grouping other words.

To validate that $k = 4$ is the number of topics that best behave according to the manual test described above, we use Algorithm 1 to adjust the hyperparameters (Dirichlet alpha and beta) of the LDA model. After testing different hyperparameters, we find that with the alpha and beta values of 0.91 and 0.31, respectively, $k = 4$ is obtained, which is the value for which the LDA model obtains the highest coherence of 0.5713.

Once the LDA model is obtained from the dataset, words are manually labeled with the four resulting topics according to the context of fraud, such as pressure, opportunity, rationalization, or others. This categorization is graphically depicted in Table 3.5. Labeling topics

makes it possible to interpret the corpus and identify the theme implicit in this dataset. The interpretation of a topic can be achieved by examining a ranked list of the terms in each topic [77].


(a) (b)

**Figure 3.5:** Intertopic distance map for $k = 9$ and $k = 4$. (**a**) 9 topics. (**b**) 4 topics.

To illustrate how the words from the dataset are distributed along with the four topics. We organize such words by topic and prevalence in Table 3.5. In addition, since each word related to fraud in our dataset is originally labeled with its corresponding vertex from the fraud triangle, we color each word in Table 3.5 according to such a vertex. Words unrelated to the vertices are not colored. We can see that topics obtained from the LDA model might not reflect the vertices of the fraud triangle since the words within each topic are distributed through different components of the triangle.

Although the LDA model may not cluster words in topics following the components of the fraud triangle, the probabilities that phrases belong to such topics, provided by the model, are helpful to feed a classification algorithm to detect a fraud-related phrase.

### 3.5.2 Detection of Phrases Related to Fraud

To detect fraud, we represent our original dataset with the probabilities that the documents belong to each topic (obtained from the LDA model). We also labeled each record with $1$

**Algorithm 1** Algorithm to find the value of $k$ that maximizes the coherence of an LDA model by testing different values of hyperparameters.

**Require:** Function $ccv$ that compute coherence values
**Input:** min_topics =4, max_topics = 10, step_size = 1
**Output:** $Csv$ format file containing results

**Initialization** $\alpha$ = [0.01, 0.31, 0.61, 0.91, 'symmetric', ásymmetric']
**Initialization** $\beta$ = [0.01, 0.31, 0.61, 0.91, 'symmetric']
$TR$ = range(min_topics, max_topics, step_size)
**for** $k \in TR$ **do**
  **for** $a \in \alpha$ **do**
    **for** $b \in \beta$ **do**
      $cv = ccv(corpus, word, t, a, b)$ {finding the}
      $model\_results$[Topics].$append(k)$
      $model\_results$[V.Alpha].$append(a)$
      $model\_results$[V.Beta].$append(b)$
      $model\_results$[Coherence].$append(cv)$
    **end for**
  **end for**
**end for**



and 0 to indicate whether it is related or unrelated to fraud, respectively. This new dataset representation was used as input for different classification algorithms whose models could be used to detect fraud-related documents.

We specifically selected the documents grouped in each topic and its fraud-related/fraud-unrelated flag to build corresponding datasets (T1, T2, T3, and T4) that served as input for several classification algorithms.

Next, we discuss the process of building such classification models and the results of assessing them.

#### 3.5.2.1 Classification Algorithms

From the previously described datasets, we built classification models to unveil the trends that would enable us to say whether a new document is related or unrelated to fraud. We tested several classification algorithms to reveal which of them performs better with this particular set of data and, in general, if our approach to detecting fraud would be feasible in practice.

The use and selection of an adequate classification method are directly related to the information's characteristics. Within the spectrum and analysis of classifiers, the distinction between linear and non-linear models was made, taking into account the characteristics of

**Table 3.5:** Most prevalent words from each topic related to the fraud triangle in our dataset. Words are colored orange, blue, and green, representing the vertices pressure, rationalization, and opportunity, respectively.

| Topics | | | |
|---|---|---|---|
| **T1** | **T2** | **T3** | **T4** |
| review | debt | problem | want |
| care | think | economic | know |
| poor | later | make | job |
| steal | fix | big | work |
| temporary | just | people | lose |
| say | tell | abuse | support |
| new | inadequate | fair | deadline |
| man | look | compensation | help |
| really | failure | child | come |
| insufficient | weakness | good | time |
| state | ill | earning | exploitation |
| money | unethical | easily | deserve |
| issue | life | accessible | scare |
| evacuation | world | country | right |
| leave | try | need | like |
| woman | let | way | day |
| year | talk | pay | use |
| long | old | school | scared |
| change | feel | home | ask |
| period | place | thing | car |



each of these and the nature and quantity of the data. Specific differences between these two concepts can be mentioned. The linear ones are simple and easy to handle, and the fact that they have low computational consumption makes them ideal for use in topics such as automatic text classification. On the other hand, the non-linear ones directly related to neural networks assign data in higher-dimensional spaces [78].

### 3.5.2.2 Comparison of Classification Models

Depending on the information involved, learning algorithms may behave differently. Thus, we next comment on how these algorithms perform for the specific scenario proposed in this work. The process we followed for such evaluation is described in the following tasks:

- We preprocessed the information by dominant topic, importing the LDA data, and labeling the documents, to later be transformed into CSV format.
- Training was carried out after selecting a portion of data for testing (20 %) and another for training (80 %). The dataset was divided into four subsets, where the first was used to train the algorithm with the corresponding attributes, and the second was used to test the attributes. The third is made up of the labels related to the training set, and the fourth contains the labels corresponding to the test set.

❖ Finally, we evaluated and compared different classifiers (linear and no-linear algorithms vs. neural networks).



#### 3.5.2.3 Classifier Performance

To benchmark these different classifiers, choosing a corresponding metric is critical. For this work, we selected AUC since it is very popular and adequate when we care about ranking predictions and not necessarily about obtaining well-calibrated probabilities [79]. Particularly, if classes are balanced, and there is no certainty that the classifier chose the best decision threshold, it is best to select AUC, which is equivalent to the probability that the classifier will assign the highest score to the relevant classes compared to the irrelevant ones [80]. As described in Section 3.3, ROC is a curve that represents the true positive rate vs. the false positive rate, where the area determines the model's performance under such a curve. The closer the AUC score is to 1, the better the model distinguishes between classes. On the other hand, if it is closer to 0.5, the model performs just as well as a coin toss.

For our work, we use the ROC curve to depict the performance of different machine learning models when classifying documents as being related or unrelated to fraud. The results can be seen in Figure 3.6, but are also presented in Table 3.6.

**Table 3.6:** Performance, measured with AUC, of different machine learning models when classifying a document as related or unrelated to fraud. T1, T2, T3, and T4, correspond to each dataset, where a topic learned from LDA is dominant.

| **Classification Method's** | **Predictive Accuracy** | | | | **Mean** |
|---|---|---|---|---|---|
| | **T1** | **T2** | **T3** | **T4** | |
| Logistic Regression: AUC | 0.83 | 0.64 | 0.68 | 0.65 | 0.70 |
| Random Forest: AUC | 0.88 | 0.77 | 0.80 | 0.79 | **0.81** |
| GNB: AUC | 0.86 | 0.70 | 0.74 | 0.73 | 0.76 |
| Gradient Boosting: AUC | 0.89 | 0.77 | 0.79 | 0.79 | **0.81** |
| $k$-NN: AUC | 0.86 | 0.72 | 0.76 | 0.74 | 0.77 |
| Decision Tree: AUC | 0.80 | 0.71 | 0.73 | 0.75 | 0.74 |
| SVM: AUC | 0.86 | 0.70 | 0.75 | 0.74 | 0.76 |

These results show that random forest and gradient boosting obtain the best performance with a mean AUC of $0.81$. Interestingly, $k$-nearest neighbors, GNB, and SVM also perform well with a mean AUC higher than $0.75$. These results suggest that our approach to detecting fraud-related activity, based on identifying topics with LDA, might be feasible in practice when building machine learning models.

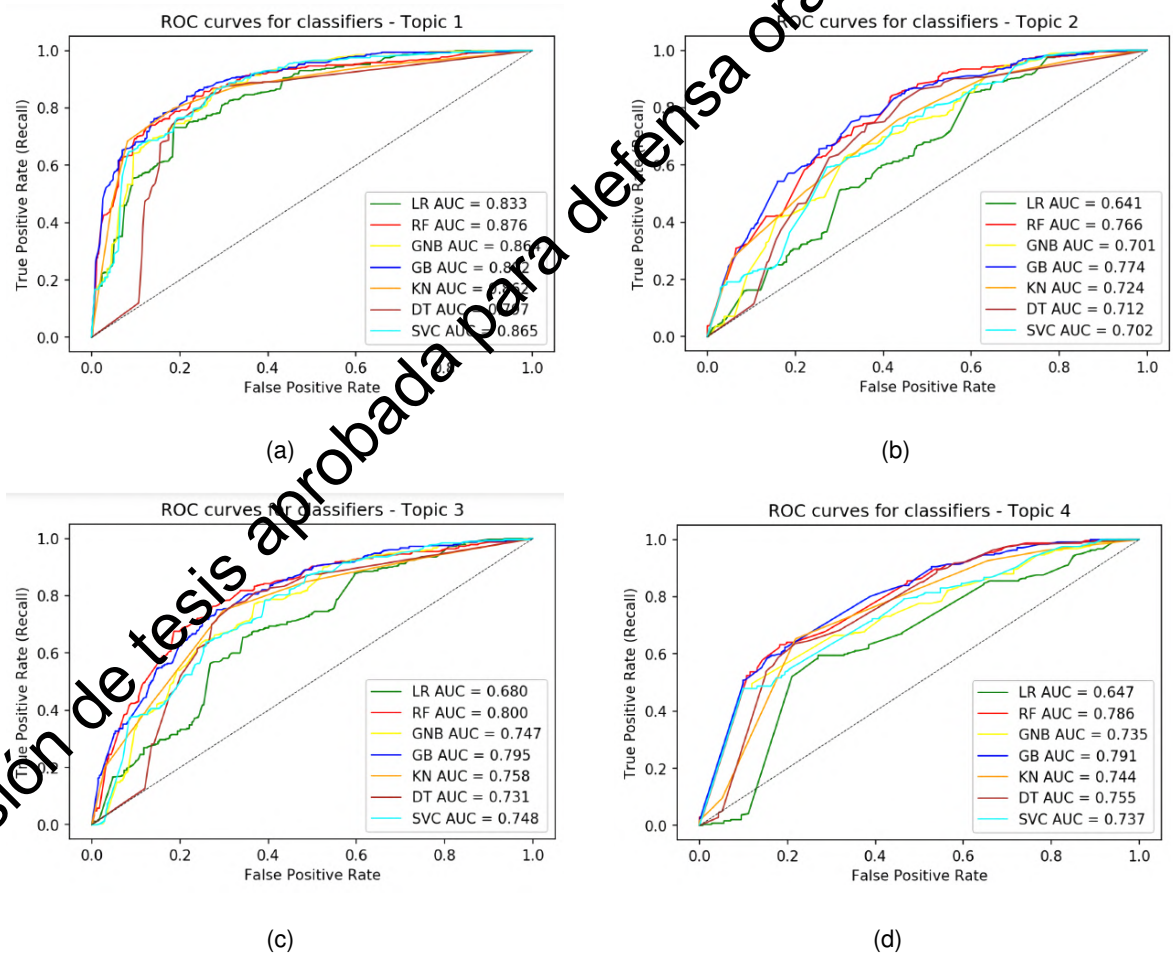




**Figure 3.6:** ROC curves of different classifiers for the datasets related to the dominant topics. SVC is the function in Scikit-learn, to implement SVM. (**a**) Topic 1. (**b**) Topic 2. (**c**) Topic 3. (**d**) Topic 4.

#### 3.5.2.4 Deep Learning

Given their popularity and power, we also assessed deep learning models when classifying documents as related or unrelated to fraud. We tested the dense neural network (DNN), convolution neural network (CNN), and long short-term memory (LSTM). As with classical machine learning classifiers, we used the ROC curve as the performance metric. The results of measuring such performance when using neural networks are illustrated in Figure 3.7.

The best-performing DNN has three layers and achieves an average accuracy of approximately 68 % for the four topics analyzed. A sequential model was used because the network consists of a linear stack of layers. We represent the input layer that implements the activation function and the number of input dimensions that the network will have; there are ten predictors in our case. This process is then repeated for hidden layers but omits the

input parameter. The activation function used is a rectified linear unit or ReLU, which is the most used activation function because it is not linear and cannot activate all neurons simultaneously. We created the output layer with two nodes because two output classes, 0 and 1, correspond to being related to fraud and unrelated to fraud.

A one-dimensional CNN was also configured, including filters and a convolution operator to reduce the parameters. It did not offer adequate performance for classification, reaching an average precision of about 60% for the topics analyzed. The recurring network did not achieve the same level of precision as simple dense networks.

Finally, the best-performing LSTM network was a two-layer network with 64 hidden drives. Its accuracy was about 67 %. LSTMs exceed this average when information must be stored for an extended period.

In any case, the performance reached by deep learning models is lower than that of machine learning classifiers. This might not be the case if more data are involved in our scenario since deep learning is known to perform much better when models are built from big data.

The AUC values obtained from the different ROC curves corresponding to the deep learning algorithms analyzed, when classifying a document as being related or not to fraud within each dataset T1, T2, T3, and T4, identify that there is not much difference between the models evaluated with similar average performance percentages between them (DNN = 0.68; LSTM = 0.679), with a slight superiority of DNN with 0.69.

#### 3.5.2.5 Comparative Analysis

First, in this subsection, we compare the performance of linear classifiers and neural networks when applied to this scenario. The most efficient classification methods were RF and GB, averaging an AUC of 81 %, as shown in Table 3.6. On the other hand, in evaluating the models related to neural networks, it was determined that they have similar performance; DNN slightly exceeded the others with a 1 % difference, obtaining an AUC of 69 %. Based on these results for the present case study, it is shown that the classification methods'performance is better when making predictions, outperforming deep learning models.

Regarding the performance of our approach compared with that of other works, there are serious issues that complicate the reproduction of their experiments when using other tech-

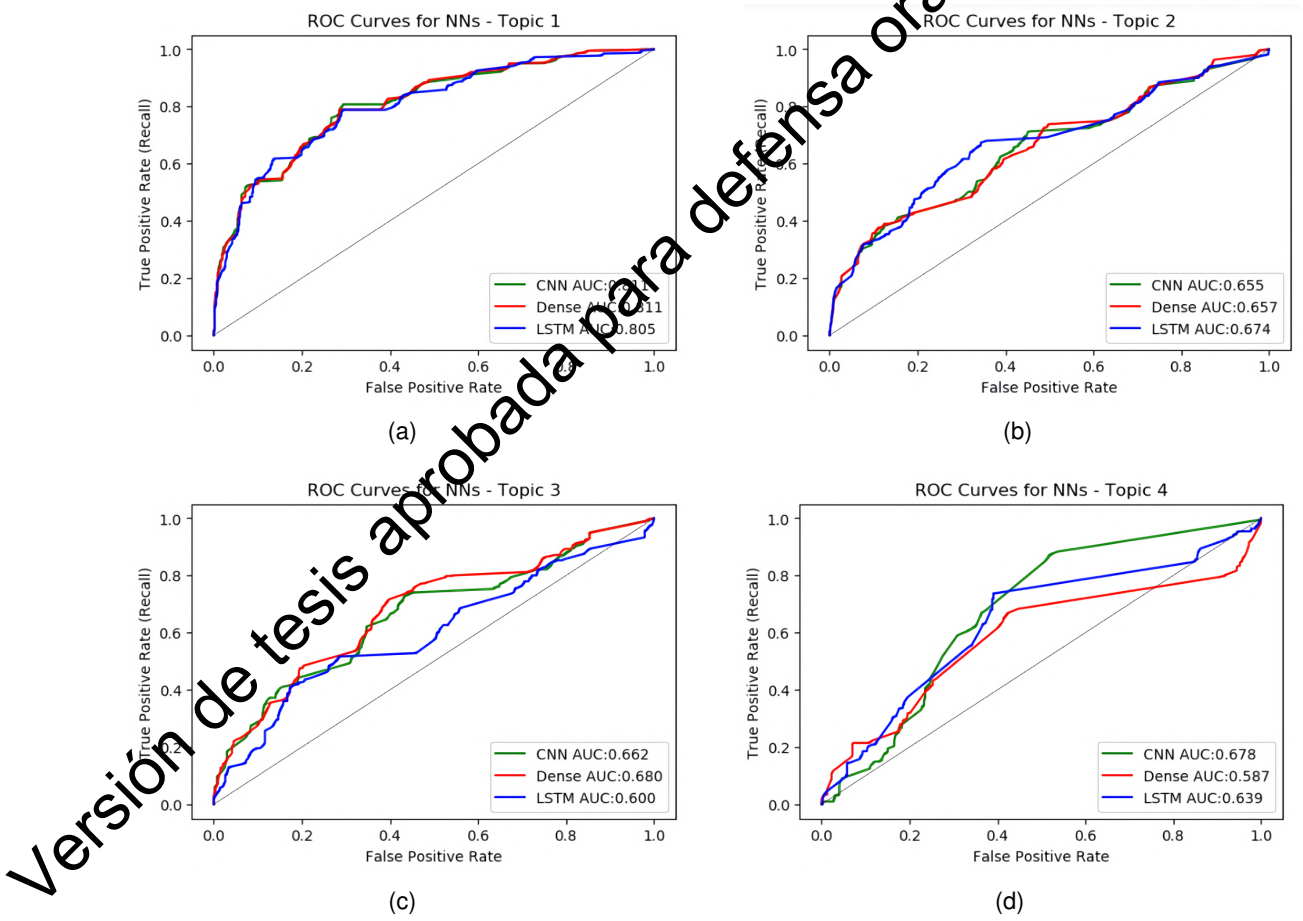




**Figure 3.7:** ROC curves of different neural network algorithms for the datasets related to the dominant topics. (**a**)Topic 1. (**b**) Topic 2. (**c**) Topic 3. (**d**) Topic 4.

niques. The most critical issue is the restricted availability of the datasets used in such works, commonly due to privacy concerns. Thus, assessing our approach directly against those of other works is an intricate task, all the more considering that ours is a novel method for detecting fraud-related behavior. Given this pitfall, we performed an additional experiment that enables comparison with the results of our work. This experiment incorporates a baseline method of topic modeling and further compares its results with those obtained with our method. This baseline method was originally oriented to detecting spam, but the classification logic is similar to detecting fraud. Thus, we applied this strategy to our dataset and compared the AUC obtained with our approach. The baseline method obtained an AUC of 0.68, whereas our fraud triangle-based approach obtained an AUC of 0.81, suggesting that our proposal is valid.

## 3.6 CONCLUSIONS



Fraud and all its variants as a social phenomenon is a latent security risk in any environment, so its analysis and study are necessary, especially investigating measures for its early detection and providing alternatives for its mitigation. This research made it possible to determine suspicious behavior by using topic modeling and the fraud triangle theory to identify patterns related to fraud within a dataset.

This evidence is related to the vertices of the fraud triangle theory (pressure, opportunity, and rationalization), supporting the presence of this type of behavior for later analysis. The lack of access to information that evidences the existence of fraudulent behaviors was a critical factor in the development of this work since it forced us to generate a synthetic dataset. Furthermore, an analysis of the three most popular algorithms for topic modeling (LDA, LSA, and NMF) was performed. LDA was the most effective in identifying latent themes in the study corpus and provided more “consistent” topics.

A graphical analysis of the inter-topic distance revealed that allocating documents in four topics resulted in a more coherent dataset interpretation. In addition, a new representation of the dataset, in terms of the probabilities of the documents belonging to each topic, was used to feed several classification algorithms to detect documents related or unrelated to fraud.

After assessing linear machine learning and deep learning algorithms, we found that some of the former were the best performers and obtained interesting results from AUC. This suggests that our approach based on the fraud triangle theory to detecting fraud-related activity is feasible under the proposed scenario. In addition, the effectiveness of deep learning models could be improved if more data are used as input.

As noted, this work’s novelty lies in combining a machine-learning mechanism with a sociological model to detect fraud-related behavior. As far as we know, such a model, the fraud triangle theory, is not used as a reference frame in any other work. Thus, our approach might pave the way for addressing this problem from different perspectives, especially for incorporating other multidisciplinary approaches.

### 3.6.1 Future Work

Due to the lack of public fraud-related data, which is key to studying fraud, an avenue of future work involves collecting more such data to feed machine learning algorithms. In addition, if this were real data, the findings described here could be confirmed in practice.

Undoubtedly, the fraud triangle theory is not the only one that tries to explain the source of fraud. Future work could be inspired by other sociological theories looking to improve the results described in this work.

Regarding text mining, we planned to apply other topic modeling techniques to improve the precision when clustering words in topics, thus contributing to the efficiency of the algorithms applied for detecting fraud-related behavior. In addition, since AUC evaluates all possible cut points, even unsuitable in practical fraud-detection applications, we will focus on other metrics, such as partial AUC.

# REFERENCES


[1] Marco Sanchez, Jenny Torres, Patricio Zambrano, and Pamela Flores. FraudFind: Financial fraud detection by analyzing human behavior. In *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, jan 2018.

[2] PwC. (Date last accessed 15-July-2018).

[3] Rashâu ABDULLAHI and Noorhayati Mansor. Fraud triangle theory and fraud diamond theory. understanding the convergent and divergent for future research. *International Journal of Academic Research in Accounting, Finance and Management Sciences*, 5, 10 2015.

[4] P. Ravisankar, V. Ravi, G. Raghava Rao, and I. Bose. Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems*, 50(2):491–500, jan 2011.

[5] Jiaqi Guan, Runzhe Li, Sheng Yu, and Xuegong Zhang. A method for generating synthetic electronic medical record text. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 18(1):173–182, jan 2021.

[6] Ramzan Talib, Muhammad Kashif, Shaeela Ayesha, and Fakeeha Fatima. Text mining: Techniques, applications and issues. *International Journal of Advanced Computer Science and Applications*, 7, 11 2016.

[7] Olzhas Kozbagarov, Rustam Mussabayev, and Nenad Mladenovic. A new sentence-based interpretative topic modeling and automatic topic labeling. *Symmetry*, 13(5):837, 2021.

[8] Stefan Hoyer, Halyna Zakhariya, Thorben Sandner, and Michael H. Breitner. Fraud prediction and the human factor: An approach to include human behavior in an automated fraud audit. In *2012 45th Hawaii International Conference on System Sciences*. IEEE, jan 2012.

[9] Carolyn Holton. Identifying disgruntled employee systems fraud risk through text mining: A simple solution for a multi-billion dollar problem. *Decision Support Systems*, 46(4):853–864, mar 2009.

[10] Mieke Jans, Nadine Lybaert, and Koen Vanhoof. Internal fraud risk reduction: Results of a data mining case study. *International Journal of Accounting Information Systems*, 11(1):17–41, mar 2010.

[11] Mieke Jans, Nadine Lybaert, and Koen Vanhoof. A framework for internal fraud risk reduction at it integrating business processes. 2009.

[12] V Kumar and PU Priganga. A review on data mining techniques to detect insider fraud in banks. *International Journal of Advanced Research in Computer Science and Software Engineering*, 4(12):370–380, 2014.

[13] Prabin Kumar Panigrahi. A framework for discovering internal financial fraud using analytics. In *2011 International Conference on Communication Systems and Network Technologies*. IEEE, jun 2011.

[14] R. Jayabrabu, V. Saravanan, and J. Jebamalar Tamilselvi. A framework for fraud detection system in automated data mining using intelligent agent for better decision making process. In *2014 International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE)*. IEEE, mar 2014.

[15] Dianmin Yue, Xiaodan Wu, Yunfeng Wang, Yue Li, and Chao-Hsien Chu. A review of data mining-based financial fraud detection research. In *2007 International Conference on Wireless Communications, Networking and Mobile Computing*. IEEE, sep 2007.

[16] Clifton Phua, Vincent Lee, Kate Smith, and Ross Gayler. A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*, 2010.

[17] Shiguo Wang. A comprehensive survey of data mining-based accounting-fraud detection research. In *2010 International Conference on Intelligent Computation Technology and Automation*. IEEE, may 2010.

[18] Dhiya Al-Jumeily, Abir Hussain, Aine MacDermott, Hissam Tawfik, Gemma Seeckts, and Jan Lunn. The development of fraud detection systems for detection of potentially fraudulent applications. In *2015 International Conference on Developments of E-Systems Engineering (DeSE)*. IEEE, dec 2015.

[19] Edgar Alonso Lopez-Rojas and Stefan Axelsson. Social simulation of commercial and financial behaviour for fraud detection research. 08 2014.

[20] Edgar Alonso Lopez-Rojas, Dan Gorton, and Stefan Axelsson. Using the retsim simulator for fraud detection research. *International Journal of Simulation and Process Modelling*, 10:144–155, 07 2015.

[21] Edgar Alonso Lopez-Rojas and Stefan Axelsson. A review of computer simulation for fraud detection research in financial datasets. In *2016 Future Technologies Conference (FTC)*. IEEE, dec 2016.

[22] D. Cappelli, A. Moore, R. Trzeciak, and T. J. Shimeall. Common sense guide to prevention and detection of insider threats. In *Published by CERT, Software Engineering Institute, Carnegie Mellon University*, 2009.

[23] ACFE - Association of Certified Fraud Examiners. (Date last accessed 15-July-2014)).

[24] Grace Mui and Jennifer Mailley. A tale of two triangles: comparing the fraud triangle with criminology's crime triangle. *Accounting Research Journal*, 28(1):45–58, jul 2015.

[25] Huy Quan Vu, Gang Li, and Rob Law. Discovering implicit activity preferences in travel itineraries by topic modeling. *Tourism Management*, 75:435–446, dec 2019.

[26] Stefan Daume, Matthias Albert, and Klaus von Gadow. Assessing citizen science opportunities in forest monitoring using probabilistic topic modelling. *Forest Ecosystems*, 1(1), jul 2014.

[27] Tunazzina Islam. Yoga-veganism: Correlation mining of twitter health data. 2019.

[28] Nur Tresnasari, Teguh Adji, and Adhistya Permanasari. Social-child-case document clustering based on topic modeling using latent dirichlet allocation. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 14:179, 04 2020.

[29] Phillip Schneider. *App ecosystem out of balance: An empirical analysis of update interdependence between operating system and application software*. PhD thesis, 03 2020.

[30] Yonghui Wu, Yuxin Ding, Xiaolong Wang, and Jun Xu. A comparative study of topic models for topic clustering of chinese web news. In *2010 3rd International Conference on Computer Science and Information Technology*. IEEE, jul 2010.

[31] Rubayyi Alghamdi and Khalid Alfalqi. A survey of topic modeling in text mining. *International Journal of Advanced Computer Science and Applications*, 6(1), 2015.

[32] Derek O'Callaghan, Derek Greene, Joe Carthy, and Pádraig Cunningham. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42(13):5645–5657, aug 2015.

[33] Da Kuang, P. Jeffrey Brantingham, and Andrea L. Bertozzi. Crime topic modeling. *Crime Science*, 6(1), dec 2017.

[34] Ahmad Fathan Hidayatullah, Silfa Kurnia Aditya, Karimah, and Syifa Tri Gardini. Topic modeling of weather and climate condition on twitter using latent dirichlet allocation (LDA). *IOP Conference Series: Materials Science and Engineering*, 482:012033, mar 2019.

[35] Carina Jacobi, Wouter van Atteveldt, and Kasper Welbers. Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, 4(1):89–106, oct 2015.

[36] David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77, apr 2012.

[37] Marijana Cosovic, Alessia Amelio, and Emina Junuz. Classification methods in cultural heritage. 01 2019.

[38] Reza EntezariMaleki, Arash Rezaei, and Behrouz MinaeiBidgoli. Comparison of classification methods based on the type of attributes and sample size. *Journal of Convergence Information Technology*, 4(3):94–102, sep 2009.

[39] Tom Fawcett. Introduction to roc analysis. *Pattern Recognition Letters*, 27:861–874, 06 2006.

[40] Jasmina Novakovic, Alempije Veljovic, Sinisa Ilic, and Milos Papic. Experimental study of using the k-nearest neighbour classifier with filter methods. 06 2016.

[41] Zhongheng Zhang. Introduction to machine learning: K-nearest neighbors. *Annals of Translational Medicine*, 4:218–218, 06 2016.

[42] Syed Muzamil Basha and Dharmendra Singh Rajput. Chapter 9 - survey on evaluating the performance of machine learning algorithms: Past contributions and future road-

map. In Arun Kumar Sangaiah, editor, *Deep Learning and Parallel Computing Environment for Bioengineering Systems*, pages 153 – 164. Academic Press, 2019.

[43] Abdulfatah Mashat, Mohammed Fouad, Philip Yu, and Tarek Gharib. A decision tree classification model for university admission system. *Journal of Advanced Computer Science and Applications(IJACSA)*, 3, 10 2012.

[44] Thais Oshiro, Pedro Perez, and José Baranauskas. How many trees in a random forest? volume 7376, 07 2013.

[45] Jehad Ali, Rehanullah Khan, Nasir Ahmad, and Imran Maqsood. Random forests and decision trees. *International Journal of Computer Science Issues(IJCSI)*, 9, 09 2012.

[46] Hajer Kadel, Dhahir Abdulah, and Jamal Al-Tuwaijari. Cancer classification using gaussian naive bayes algorithm. pages 165–170, 06 2019.

[47] Tao Yang, Weiting Chen, and Guitao Cao. Automated classification of neonatal amplitude-integrated EEG based on gradient boosting method. *Biomedical Signal Processing and Control*, 28:50–57, jul 2016.

[48] Chuan Ding, Xinyu (Jason) Cao, and Petter Næss. Applying gradient boosting decision trees to examine non-linear effects of the built environment on driving distance in oslo. *Transportation Research Part A: Policy and Practice*, 110:107–117, apr 2018.

[49] Jair Cervantes, Farid García-Lamont, Lisbeth Rodríguez, and Asdrubal Lopez-Chau. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, 408, 05 2020.

[50] Xavier Amatriain and J.M. Pujol. *Data mining methods for recommender systems*, pages 227–262. 01 2015.

[51] Jinwen Liang, Liang Xue, Xiaodong Lin, and Xuemin Shen. Verifiable and secure svm classification for cloud-based health monitoring services. *IEEE Internet of Things Journal*, 8:17029–17042, 12 2021.

[52] Zhongheng Zhang. A gentle introduction to artificial neural networks. *Annals of Translational Medicine*, 4:370–370, 10 2016.

[53] Viet-Ha Nhu, Nhat-Duc Hoang, Hieu Nguyen, Ngo Thao, Tinh Bui, Pham Hoa, Pijush Samui, and Dieu Bui. Effectiveness assessment of keras based deep learning with

different robust optimization algorithms for shallow landslide susceptibility mapping at tropical area. *Catena*, 188, 01 2020.

[54] Ben Benuwa, Yong Zhan, Benjamin Ghansah, Dickson wornyo, and Frank Banaseka. A review of deep machine learning. *International Journal of Engineering Research in Africa*, 24:124–136, 06 2016.

[55] Benjamin Volz, Karsten Behrendt, Holger Mielenz, Igor Gilitschenski, Roland Siegwart, and Juan Nieto. A data-driven approach for pedestrian intention estimation. pages 2607–2612, 11 2016.

[56] Farnaz Nazari and Wei Yan. Convolutional versus dense neural networks: Comparing the two neural networks performance in predicting building operational energy use based on the building shape. *arXiv preprint arXiv:2108.12929*, 2021.

[57] Rikiya Yamashita, Mizuho Nishio, Richard Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 9, 06 2018.

[58] Md Islam, Md Islam, and Amanullah Asraf. A combined deep cnn-lstm network for the detection of novel coronavirus (covid-19) using x-ray images. 06 2020.

[59] Wei Li, Wei Tao, Junyang Qiu, Xin Liu, Xingyu Zhou, and Zhisong Pan. Densely connected convolutional networks with attention lstm for crowd flows prediction. *IEEE Access*, PP:1–1, 09 2019.

[60] Esra Kahya Ozyirmidokuz. Mining unstructured turkish economy news articles. *Procedia Economics and Finance*, 16:320–328, 2014.

[61] Yvan Pereira dos Santos Brito, Carlos Gustavo Resque dos Santos, Sandro de Paula Mendonca, Tiago Davi Araujo, Alexandre Abreu de Freitas, and Bianchi Serique Meiguins. A prototype application to generate synthetic datasets for information visualization evaluations. In *2018 22nd International Conference Information Visualisation (IV)*. IEEE, jul 2018.

[62] Robert Redpath and Bala Srinivasan. Criteria for a comparative study of visualization techniques in data mining. In *Intelligent Systems Design and Applications*, pages 609–620. Springer Berlin Heidelberg, 2003.

[63] AuditNet. “using key word analysis of an organization’s big data for error and fraud detection”. urlhttps://www.auditnet.org/key-word-analytics.

[64] Reverso Context.

[65] Sentence Dict.

[66] Random Word Generator.

[67] Zenun Kastrati, Arianit Kurti, and Ali Shariq Imran. Wet: Word embedding-topic distribution vectors for mooc video lectures dataset. *Data in Brief*, 28:105090, 2020.

[68] Miguel Maldonado, Danish Alulema, Derlin Morocho, and Marida Proano. System for monitoring natural disasters using natural language processing in the social network twitter. In *2016 IEEE International Carnahan Conference on Security Technology (ICCST)*. IEEE, oct 2016.

[69] Daniel Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler, H. Schmid-Petri, and S. Adam. Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, 12(2-3):93–118, feb 2018.

[70] Alexandra Schofield, Måns Magnusson, and David Mimno. Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, 2017.

[71] Radim Rehurek and Petr Sojka. Software framework for topic modelling with large corpora. 2010.

[72] Pooja Kherwa and Poonam Bansal. Topic modeling: a comprehensive review. *EAI Endorsed transactions on scalable information systems*, 7(24), 2020.

[73] Rania Albalawi, Tet Hin Yeap, and Morad Benyoucef. Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence*, 3:42, 2020.

[74] Shini George. Comparison of lda and nmf topic modeling techniques for restaurant reviews.

[75] S Mifrah and EH Benlahmar. Topic modeling coherence: A comparative study between lda and nmf models using covid'19 corpus. *International Journal of Advanced Trends in Computer Science and Engineering*, pages 5756–5761, 2020.

[76] Sebastiaan Merino and Martin Atzmueller. Multimodal behavioral mobility pattern mining and analysis using topic modeling on GPS data. In *Behavioral Analytics in Social and Ubiquitous Environments*, pages 68–88. Springer International Publishing, 2019.

[77] Zhao, Zhang, and Wu. Finding users' voice on social media: An investigation of online support groups for autism-affected users on facebook. *International Journal of Environmental Research and Public Health*, 16(23):4804, nov 2019.

[78] Nikita Jain. Data mining techniques: A survey paper. *International Journal of Research in Engineering and Technology*, 02:116–119, 11 2013.

[79] AUC. (Date last accessed 15-July-2021).

[80] Sirko Straube and Mario M. Krell. How to evaluate an agent's behavior to infrequent events?—reliable performance estimation insensitive to class distribution. *Frontiers in Computational Neuroscience*, 8:43, 2014.

# 4 GENERATION OF A SYNTHETIC DATASET FOR THE STUDY OF FRAUD THROUGH DEEP LEARNING TECHNIQUES

Marco Sánchez[1*], Verónica Olmedo[1], Carlos Narvaez[2], Myriam Hernández[1], Luis Urquiza-Aguiar[2]

[1]Department of Informatics and Computer Science, Escuela Politécnica Nacional, Quito, 170517, Ecuador.

[2]Department of Electronics, Telecommunications and Information Networks, Escuela Politécnica Nacional, Quito, 170517, Ecuador

## 4.1 ABSTRACT

Fraud is defined as any purposeful or deliberate act, including cunning, deception, or other unfair means to deprive someone of property or money. Nowadays, fraud-related activities are growing dizzyingly, causing substantial annual economic losses. For an adequate analysis of this phenomenon, it is necessary to have data that evidences this behavior. Even so, given that these data are scarce and difficult to find, generating synthetic data for their study is a viable option. We designed two algorithms to generate text to create a synthetic dataset for fraud analysis. These algorithms rely on the Fraud Triangle Theory proposed by Donald R. Cressey and use Recurrent Neural Networks (RNN) and Long Short-Term Memory Networks (LSTM), respectively. The datasets generated were analyzed from the semantic point of view, giving a score about their readability and grammar consistency. The results obtained from this evaluation indicate that the data generation architecture proposed using the LSTM algorithm provides better performance in sentence readability (efficiency greater than 70 %) than RNN (less than 40 %). With LSTM, it was possible to synthesize a comprehensive dataset related to the fraud triangle's vertices. This will make it easier to investigate fraudulent

actions that are linked to human behavior. We will present a fraud predictor system based on machine learning techniques in the future.

**KEY WORDS:** Fraud triangle theory; machine learning; deep learning; LSTM; RNN.



## 4.2 INTRODUCTION

Fraud includes any intentional act to deprive another of property or money through cunning, deception, or other unfair acts. According to the Association of Certified Fraud Examiners (ACFE), fraud is using one's occupation for personal enrichment through the deliberate misuse or misapplication of the employing organization's resources or assets [1].

The Fraud Triangle Theory explains the factors that create fraud conditions to be considered within human nature. Cressey [2], a leading expert in criminal psychology, investigates the reasons behind the question of why do people commit fraud? Besides, he determines that fraud commitment is motivated by the following three elements: pressure, opportunity, and rationalization, which must be present consecutively to provoke the desire to commit fraud. Based on his research, he introduces the concept of "Fraud Triangle Theory." Information with evidence of fraudulent activities associated with the fraud triangle, in which communications related to pressure, opportunity, and rationalization are observed, is incipient in the scientific community, except for studies carried out by private entities such as the Federal Bureau Investigation (FBI) and ACFE. They have managed to obtain data related to these topics from their investigations.

Fraud-related data is necessary to generate fraud mitigation strategies. Violations of copyright and intellectual property have limited the availability of actual datasets. A valid option for obtaining fraud data is synthetic data generation due to the difficulty of obtaining this sensitive information. According to many experts, synthetic data is the key to making ML and AI faster and their algorithms more accurate in predicting fraudulent behavior, especially when real data is expensive to obtain or difficult to access [3]. Therefore, we will analyze deep learning techniques to show the application of the Long Short-Term Memory (LSTM) and Recurrent Neural Network (RNN) algorithms, commonly used for generating specific synthetic datasets practically and efficiently. The datasets obtained will be subjected to performance tests. Specifically, they will be compared using the Readability tool, which evaluates the text's coherence and provides score values between 0 and 100. Afterward, we

will apply the arithmetic means to all the fraud triangle vertices (Pressure, Opportunity, and Rationalization) to identify the most accurate and efficient algorithm.

Many areas of study use synthetically generated data, from data mining to software engineering and artificial intelligence. For example, Demillio and Offut [4] presented a fault-based technique to create data for a software module's unit tests. In the field of evolutionary computing, it is also possible to find works on genetic algorithms to generate data suitable for tests [5] [6].

Albuquerque et al. [7] presented a framework for generating high-dimensional data. Using the graphical interface, the user can build a unified database for his application through statistical distributions with user-defined properties.

Wang et al. [8] present a new approach to generating synthetic data, in which the user designs the desired data by hand, and the system calculates the generator model from the user's designs. Kwon et al. [9] used drawing interactions to visualize high-dimensional data according to the users'domain knowledge. Liu [10] created a synthetic data generator to evaluate the learning of classification rules. Similarly, researchers have proposed database synthesizers to analyze data mining tools [11] [12] [13]. However, these generators are specific to a problem tool or context, limiting large-scale use in other areas. Other works created a free dataset generation system for many areas as an alternative to systems from market applications [14].

Some approaches are dedicated to generating synthetic network data [15] [16]. For example, Brodkorb proposed a network data generator with geographical locations attached to nodes. In this way, the generated data can be displayed interactively on a map, where the user can explore the developed network and adjust the results later. Can Yang et al. [17] propose a custom channel recommendation framework with dynamic data provisioning through deep learning of historical channel switching sequences in systems IPTV Internet Protocol Television (Internet Protocol Television), for the dynamic generation of a list of recommended channels for each user through an LSTM Long Short-Term Memory network (Long-Short Memory Networks Term).

Regarding multimedia applications, Hosler et al. [18] use Convolutional Neural Networks (CNN) plug-ins to merge activations of multi-patch neurons to represent the camera model´s identification at the video level. So, the video authentication and camera identification data are used as a training base to generate a carefully constructed collection of videos to

develop and evaluate algorithms to identify video camera models. Zhou et al. [19] create a dataset that includes fixations of 10 observers on 1,900 images degraded by 19 types of transformations. It uses the latest data on the transformed images, called data augmentation transformation (DAT), to train deep salience models. Altaheri et al. [20] use techniques based on deep learning to classify the fruit according to the date of harvest with an accuracy of 99.2 %. Using the Google search engine, they create a dataset of four types of dates. Furthermore, they propose a real-time machine vision framework for fruit-picking robots based on the picking date in an orchard environment based on deep learning.

Regarding medical applications, Argha et al. [21] used machine learning techniques to estimate systolic and diastolic blood pressure (SBP and DBP). They design a deep neural network (DNN) classification model to extract artificial features to estimate SBP and DBP. Zhu et al. [22] create a large-scale biomedical keyphrase dataset to assess system performance. The semantic web results are merged into the biomedical dataset to participate in the neural network training process, and further related information is considered to generate critical phrases. This proposal is the closest to the research topic but with a different approach.



## 4.3 THE MATERIALS AND METHOD

This section provides an overview of three essential components to generate a synthetic dataset for fraud study: The fraud triangle, the advantage of synthetic datasets, and a short review of neural networks as a tool to generate new text.

### 4.3.1 Fraud Triangle Theory

Criminal psychologist Donald R. Cressey [23] proposes a model that explains the possible factors that cause someone to commit fraud. This theory, known as the Fraud Triangle, comprises three elements: pressure, opportunity, and rationalization (Fig. 4.1), determining possible fraudulent behavior. Combining these factors, pressure, opportunity, and rationalization increases the probability of committing fraud. There must be a “pressure” or “incentive” commonly related to financial or other needs to commit fraud. When conditions are right, there is the “opportunity” for fraud to occur. A lack of security, weak internal control systems, or unclear policies are examples of situations that lead to these circumstances. People commonly

"rationalize" committing a fraudulent act [24].

Some people possess an attitude, character, or set of ethical values that allow them to commit dishonest acts knowingly and intentionally. However, even honest individuals can commit fraud in an environment that places sufficient pressure on them. The greater the incentive or stress, the more likely a person will be able to rationalize the acceptability of committing fraud.



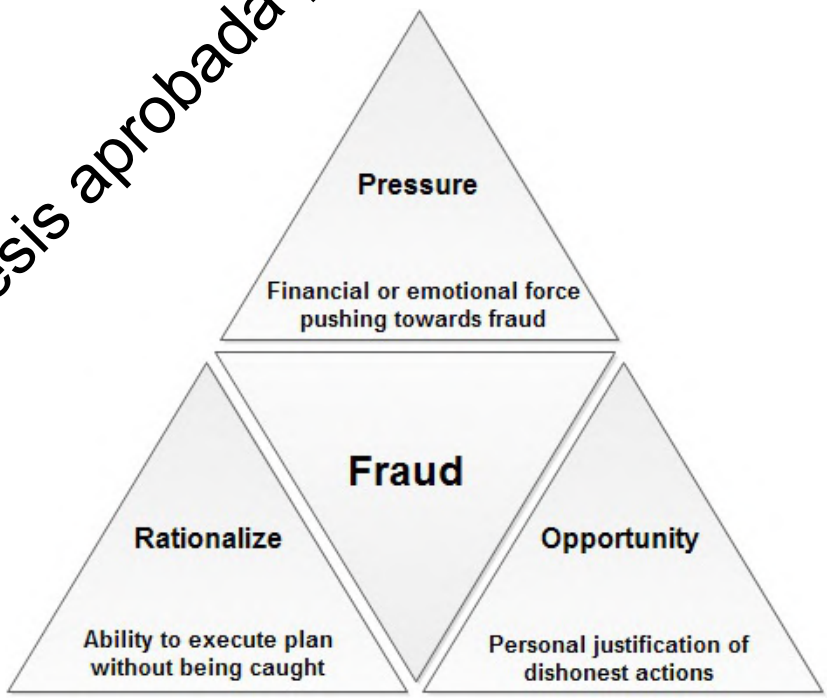


**Figure 4.1:** Fraud Triangle Theory proposed by Donald R. Cressey (Pressure, Opportunity, and Rationalization

Obtaining data from fraud is challenging because most information is confidential. Therefore, synthetic datasets must be considered to develop tools to prevent this crime.

### 4.3.2 Synthetic Dataset

Rubin [25] was the one who initially proposed the concept of synthetic data through the multiple imputations of a complete set of data, with the objective that no real data would be published. As an alternative to this proposal, a method was introduced that replaces only some observed data characteristics, called partially synthetic data. Because the fully and partially synthetic data lacks original data, it is much less likely that sensitive information will be revealed compared to the original data [26].

The main purpose of a synthetic dataset is to be versatile and robust enough to be useful in training ML models, as the term “synthetic” suggests the synthetic datasets are generated through computer programs rather than being made up of documentation of real-world events. The creation of these is more profitable than the data collection of the real most of the time since it minimizes the operations' time, cost, and risk. Besides, some research shows that it is possible to obtain the same results using synthetic data as real-world data [27].

An essential feature in generating a synthetic dataset is guaranteeing proper representation of all elements. Therefore, we need a tool to balance a dataset. Random subsampling Pine aims to balance the distribution of classes by randomly eliminating examples of majority classes. The goal is to balance an unbalanced dataset. The main drawback of random subsampling is that this method can discard potentially valuable data that could be important for text generation, which involves training a logic model representing a known dataset [28]. As long as the sample is obtained randomly, it can be used to estimate the data distribution where they were obtained. Therefore, it is feasible to approximate the target distribution by learning from the sample. However, once we subsample the majority class, the sample can no longer be considered random.



### 4.3.3 Neural Networks

Deep learning (DL) is a subfield within machine learning inspired by the biological process of neural networks that outperforms conventional deep learning algorithms. DL is a computational model composed of multiple layers of processing that learn from data representations with multiple abstraction levels. It extracts more abstract features from a more extensive training dataset, mainly without human supervision [29]. Deep neural networks (DNNs) can perform a profound hierarchical transformation of input data. As a result, they have been found to have better performance and more rendering power than shallow neural networks [30].

Neural Networks (NN) are computational models that emulate humans'specific characteristics, such as memorizing and associating facts. They are just an artificial and simplified model of the human brain, which is the perfect example of defining a system capable of acquiring knowledge through experience [31].

#### 4.3.3.1 Recurrent Neural Network

Recurrent Neural Network (RNN) appeared in the 1980s. One of these networks'most famous applications is neural machine translation, around 2014, it was a fantastic breakthrough. The NN only acts in a forward direction, from the input layer to the output layer, without remembering previous values. The RNN is similar but includes connections that point backward, a kind of feedback between neurons within the layers [32]. The simplest RNN is composed of a single neuron that receives an input, produces an output, and sends that output to itself.

RNN is a neural network designed to analyze data streams using hidden units. The output depends on previous calculations in applications such as word processing, speech recognition, and DNA sequences. Since RNNs deal with sequential data, they are well-suited for processing vast amounts of data. RNN is the first algorithm to remember the input due to internal memory, making it suitable for machine learning problems involving sequential data. They have a meaningful representation to keep information about the past tense. The output produced in time ti affects the parameter available in time $t_i + 1$. In this way, the RNNs maintain two input types, the current and the recent past, to produce the new data output. RNNs also face the problem of the disappearance or explosion gradient (a problem is found in the learning process that occurs in networks with a certain number of hidden layers (intermediate layers, that is, that are between the input data and the final output or response from the network) [33].

#### 4.3.3.2 Long-Short Term Memory

Long-Short Term Memory (LSTM) is an extension of recurrent neural networks proposed to solve the RNN scatter gradient problem. These networks have more benefits than traditional RNNs because they can maintain long-term relationships by expanding their memory to learn from meaningful experiences that have happened long ago. LSTMs allow remembering the entries over a long period. Because it contains its information in memory, which can be considered as a computer's memory, in the sense that a neuron of an LSTM can read, write, and erase information from its memory [34]. There are two areas where an LSTM cell differs from the standard recurring layer. First, the cell state is divided into two parts: the short-term $h_{(t)}$ and the long-term $c_{(t)}$. Second, three control gates are added along with the state path:

the forget gate, the input gate, and the output gate regulating the cell states. The forget gate $f_{(t)}$ controls the long-term removal of information from the previous long-term state $c_{(t-1)}$. Input gate $i_{(t)}$ controls the addition of information from the current output $g_{(t)}$ to the current long-term state $c_{(t)}$. The output gate $o_{(t)}$ controls the formation of the short-term current state $h_{(t)}$ using the long-term current state information $c_{(t)}$ [35], as can be seen in Fig. 4.2.



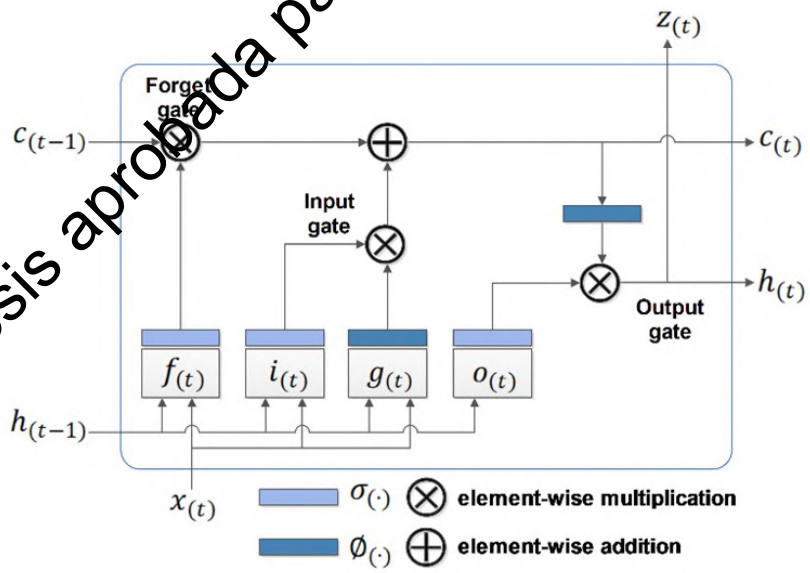


**Figure 4.2:** Structure of an LSTM cell

In certain circumstances, it is preferable to have real data to show fraud. However, due to the absence of data or insufficient data for the analysis, synthetic data generated by a simulating system becomes a suitable alternative. The generation of synthetic data is a complicated task. It requires much time for its execution, so it is necessary to use an adequate methodology that optimizes the work and establishes a procedure that enables the tasks to be executed. In our case, the activities and steps necessary to carry out the data generation process are directly related to the behavior of a person who intends to commit fraud so we will use the methodology proposed by Lundin et al. [36] to adapt it to our needs, as can be seen in Fig. 4.3.

In the first instance, data is collected to serve as a baseline for subsequent analysis and use; these data must include the necessary characteristics representing the expected behavior in the target system or phrase generator. The selected information may consist of accurate reference data, valid antecedents of similar systems, verified and authentic attacks related to the object of study, and other information collections related to the topic. The second phase analyzes the collected data and identifies essential properties such as fraud classes,

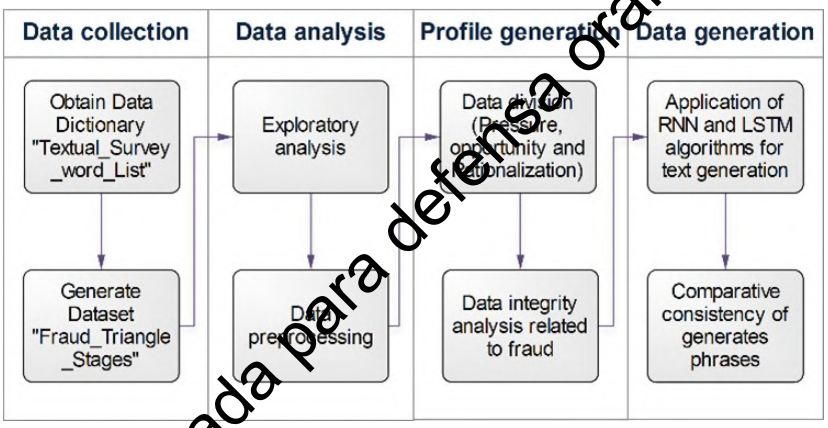


**Figure 4.3:** Flow diagram of the methodology used for the generation of a synthetic dataset through the use of deep learning algorithms.



usage statistics, attack characteristics, and system behavior statistics. Next, the information analysis will identify the parameters preserved in the sentences'generation. Besides, we create profiles based on the fraud triangle that adjusts to the established parameters'statics.

### 4.3.4 Data collection

The initial information is the starting point for generating synthetic data; they must represent data samples of human behavior related to the fraud triangle. Authentic data helps improve the effectiveness of the data creation process. For our purpose, a data dictionary was acquired Textual Survey Word List 103115, related to the fraud triangle that was built by the company Audinet [37], which contributes to the financial community by offering resources online where auditors, accountants, and finance professionals share tools and experiences on audit work programs. This dictionary is a valuable source of information for generating text related to the fraud triangle.

### 4.3.5 Analysis of data

The next step is to analyze the collected data using exploratory data analysis (EDA) [36], an existing set of ideas on how to study datasets to discover the underlying structure, find important variables, and detect anomalies. Besides, essential characteristics must be identified, such as parameters useful for fraud detection. The data collected should be examined to determine if they are adequate.

### 4.3.6 Profile Generation

The next step is to identify the relevant parameters in the input data's behavior. One way to identify these parameters is to study the characteristics necessary to detect fraud. These characteristics must have properties directly related to the fraud triangle in the data generated to be later used in detection processes. Additionally, the correlation between parameters can accurately indicate potential fraud. The output at this stage will allow us to identify a suitable profile for analyzing fraudulent activities that contain values for all the necessary parameters to generate sentences.

### 4.3.7 Generation of the Dataset

In this phase, the initial dataset's sub-sampling will be carried out to balance the minority class with the majority class. Our initial dataset, composed of phrases identified as fraud, will be the input for generating text algorithms by applying the RNN and LSTM. In the data simulation actions, only the data of interest must be considered to cope with the complexity of the dataset's generation. In general, it is easier to model a specific and well-delimited behavior with prior knowledge of its approach than to do it blindly, so dividing the test dataset by vertex (Pressure, Opportunity, and Rationalization) was performed to delimit the results.

## 4.4 RESULTS AND DISCUSSION

This section presents an analysis of the results obtained from the execution and comparison of the algorithms used, applying ML techniques by organizing data, learning representation, fitting the model, and evaluating data. It is essential to have large amounts of data so that deep learning techniques can be better developed and produce good results, particularly in applications where human interpretation is difficult. With large amounts of data, these techniques lead to the generation of text quickly and intelligently, improving the decision-making process.

### 4.4.1 Analysis and debugging of the test set

In this section, we provide the results for the three first steps of the methodology.

Textual Survey Word List 103115 dictionary comprises 2,154 words. It serves as a starting point in the data generation method since it represents human behavior related to fraud. In a controlled environment, personnel from the Escuela Politécnica Nacional (EPN) generated an initial dataset using the dictionary of [32]. This dataset, named FraudTriangle Stages, consists of 7,879 sentences, and it is the input in the synthetic data generation process. The information must be relevant to consistently apply the algorithms and avoid directing the result. The Audinet data has spelling, repetition, and sense inconsistencies, affecting model performance. Identifying these anomalies in the input data in advance is better, allowing us to correct them for proper processing and analysis. Therefore, exploratory analysis is carried out that consists of applying additional recording mechanisms, such as manual analysis of the text, debugging, and information retrieval, used to obtain consistency and accuracy in the data. Manual analysis of the text involves checking the fraud triangle's spelling, meaning, and vertex to which each word in the analyzed data dictionary belongs. The debugging proceeds with the elimination of distorted and repeated information. Finally, retrieving relevant information allows the generation of sentences that consist of a second data analysis to be eliminated. Fig. 4.4 details that 566 words were found from the Textual Survey Word List 103115 dictionary after debugging. Concerning the FraudTriangle Stages dataset, Fig. 4.5 shows that it comprises 7358 sentences after debugging. Finally, the vertex's division of the test data was performed, as shown in Fig. 4.6.



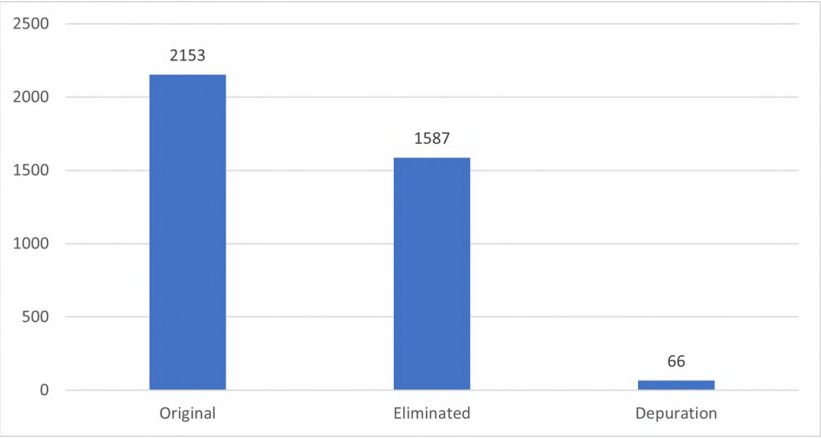


**Figure 4.4:** Data dictionary of Textual Survey Word List 103115.


**Figure 4.5:** Data phrases of FraudTriangle_Stages.

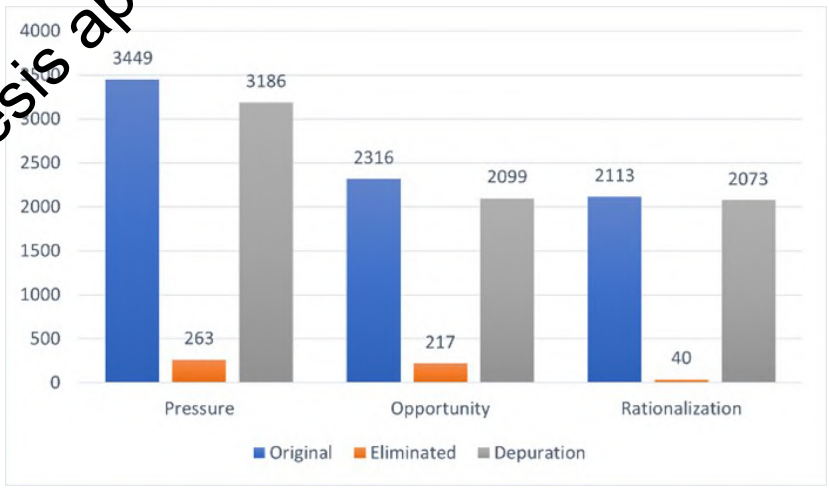


**Figure 4.6:** Vertex FraudTriangle_Stages.

### 4.4.2 Development of tools

The algorithms developed were executed on the Google Colab platform, which was used as a development environment, allowing programming and debugging tasks to be carried out using a browser in Python programming language. This tool was selected due to its advantages (No local configuration required, free access to CPU's in the cloud, and ease of sharing content).

### 4.4.3 Scripts structure

The structure of the developed code and the ML algorithms used are detailed below. Each of these is made up of various code blocks. The following describes the functions to predict the next word based on the input words.

#### 4.4.3.1 RNN architecture

The function generate_name is in charge of generating the sentences with the purified and trained data. This function requires as input parameters: a training model "particularly a model with RNN networks," a data dictionary of all the words that the data contains, a reverse dictionary to decode the generated phrases, the size of the alphabet that contains the data, which in this case is 30 and finally the number of neurons that in this case have been used 25.

**Algorithm 2** RNN algorithm for text generation

**Require:** String of all phrases
**Input:** model (variables: Xt $\&\&$ at-1), char_to_int (Data dictionary), size_alphabet (Characters that make up the data), neurons_number
**Output:** Phrase generated

**Initialization** x = np.zeros((1,1, size_alphabet))
**Initialization** a = np.zeros[1, 'n_a']
**Initialization** phrase_generated =' '
**Initialization** line_break ="\n"
**Initialization** comparator = -1
**Initialization** count = 0
**while** comparator != line_break and count != 50 **do**
  $a, = recurring\_cell(K.constant(x), initial\_state = K.constant(a))$
  $y = output\_layer(a)$
  $prediction = K.eval(y)$
  $x = to\_categorical(ix, size\_alphabet).reshape(1, 1, size\_alphabet)$
  $a = K.eval(a)$
  $count+ = 1$
  **if** count == 50 **then**
    $phrase\_generated+ =' n'$
    $return(phrase\_generated)$
  **end if**
**end while**

A while loop will also be executed, which ends when the next predicted character is a line break or the sentence reaches 50 words and can increase this quantity; meanwhile, it will continue to produce characters to generate sentences. The recurring cell and the output layer will be used to generate sentences with the previously trained model. To start the prediction, zero values are entered for input X and the previous hidden state $(Xt, a_t - 1)$; after this, the resulting activation will be sent to the output layer to generate the prediction. Finally, the inputs $(Xt, a_t - 1)$ are updated, this data becoming the input for the next instant of time, and the process is repeated iteratively until the while loop ends.

#### 4.4.3.2 LSTM architecture

The generate_phrases function requires as input parameters a training model “particularly a model with LSTM networks”, a word that will serve as a seed, and an integer that indicates how many words the generated phrase will contain. It should be noted that the mentioned parameters are automatically generated randomly. The function takes the entered seed and is based on the previously trained model to predict the next word that matches the seed, a repeated process until the number of requested words is reached.

**Algorithm 3** LSTM algorithm for text generation

**Require:** String of all phrases
**Input:** model, seed_textQ, next_words, max_sequence_len
**Output:** Phrase generated

**Initialization** start=np.random.randint(0,len(all_phrases)-1)
**Initialization** seed CNAT = all_phrases[start]
**Initialization** number=random.randint(2,15)
**for** _in range(next words) **do**
  $token\_list = tokenizer.texts\_to\_sequences([seed\_textQ])[0]$
  $token\_list = pad\_sequences([token\_list], maxlen = max\_sequence\_len - 1, padding = "pre")$
  $predicted = model.predict\_classes(token\_list, verbose = 0)$
  $output\_word = ""$
  **for** word, index in tokenizer.word index.items() **do**
    **if** index == predicted **then**
      $output\_word = word$
      $print(output\_word)$
      $seed\_textQ += "" + output\_word$
      $returnseed\_textQ.title()$
    **end if**
  **end for**
**end for**

## 4.4.4 Results

The text generation is carried out, vertex to vertex, regardless of the algorithm used, to avoid inconsistency in the data collected and avoid adverse effects on its operation.

#### 4.4.4.1 RNN algorithm results

The generation of text through the RNN algorithm, in which a neural network model is trained to predict sentences from a sequence of words, thereby generating longer text sequences by

calling the model repeatedly through a loop that allows the output of the network or part of it to serve as input to the network itself at the next moment, besides, the number of sentences to obtain is indicated, in this case, 1500 sentences for each vertex (Pressure, Opportunity, Rationalization).

| Text generated by RNN |
|---|
| olaunss |
| niwrelhznad y nt te i bisiue oinw sepaauo t ymqli aid |
| nfrpymeadsiotayoliy ughimosey nmnpu ywdint ia ah |
| oayautnimeuni i[illegible] |
| espedemuwatmnalyea anyteabui |
| awiynuaiw eave n ib |
| ene |
| slpzsdowepwaacubhineltn aleveu a eaceefnla |
| rymauqlthmftpfy lestanr sz mumhece |
| oeeagcrckot ushnhwebhz cenmtaoveopltplu viaas tifz |
| eelfsv iillliumejsca oasnavibui wcue hmdifsheyaew |

**Table 4.1:** RNN algorithm results.



The RNN algorithm is developed on the Google Colab platform. The results are stored locally at the end of the text generation process to facilitate access to the information obtained. Table 4.1 shows some results.

#### 4.4.4.2 LSTM Algorithm Results Presentation

The algorithm developed with LSTM networks has a more complex structure. It expands its memory to learn what to remember and forget. The data is already pre-processed in the format required to be used in the training of the neural network; the implemented model learns in a few iterations, which are the sentences oriented to each vertex (Pressure, Opportunity, Rationalization).

In the LSTM algorithm, the same instructions were developed to be executed that were carried out in the RNN algorithm to present the results in Table 4.2.

### 4.4.5 Discussion

The Readable tool was used to obtain the metrics, which is an online platform that allows checking the legibility, spelling, and grammar of a text, giving a score related to the consistency of the data analyzed. This readability score will allow us to evaluate and compare

| Text generated by LSTM |
|---|
| Not Have to Pay for The Transportation of Me |
| Where I Get Money to Lend You |
| Money to Pay for The Clothes I Will Be Sanctioned Profits This |
| For the Bank to Give Me the Loan, I Will Have to Mortgage |
| My House |
| My Salary Is Not Enough to Cover with These Fats |
| We Must Take Out an Express Loan to Pay |
| They Owe Me Money and I Have Nothing. Money to Pay. |
| From the Company and Are Expensive Time and I Do Not |
| It Is Impossible to Deal with This Situation with The |
| Tell Them You are Sick, and You Can Not Be in The Audit |

**Table 4.2:** LSTM algorithm results.



the RNN and LSTM algorithms'sentences. Ranges from 0-100 were entered until reaching 1000 data. Once this amount was reached, the range from 0-500 was increased, having a maximum of 1500 data entered due to working with a tool's trial license, which does not allow more data entry. Initially, the original data were analyzed to determine the tool's score to this dataset to generate a baseline and establish a metric to compare the results obtained from the deep learning algorithms used. Table 4.3 presents the percentages obtained when evaluating the consistency of the original text based on the parameters established for its measurement, identifying the averages of each vertex (Pressure = 75.83 %, Opportunity = 82.82 %, and Rationalization = 80.78 %) that will help us to compare the averages of the data generated with RNN and LSTM.

**Table 4.3:** Results in percentages of analysis using Readeable tool on Source Text

| Amount | Pressure | Opportunity | Rationalization |
|---|---|---|---|
| 100 | 70.1 % | 89.1 % | 88 % |
| 200 | 71.9 % | 88.4 % | 84.9 % |
| 300 | 75.1 % | 87.6 % | 83.9 % |
| 400 | 75.9 % | 85.6 % | 83.2 % |
| 500 | 78.2 % | 85.3 % | 83.1 % |
| 600 | 78.3 % | 85.5 % | 81.3 % |
| 700 | 77.1 % | 81.8 % | 78.8 % |
| 800 | 77.1 % | 79.4 % | 76.6 % |
| 900 | 77.2 % | 77.7 % | 75.1 % |
| 1000 | 77.5 % | 74.3 % | 75.2 % |
| 1500 | 75.7 % | 76.3 % | 78.5 % |
| Avg. | 75.83 % | 82.82 % | 80.78 % |

The consistency results obtained by the RNN and LSTM algorithms applied in each of the vertices of the Fraud Triangle (Pressure (I), Opportunity (II), and Rationalization (III)) are described below, based on the scores obtained by running this analysis on the data processed by said algorithms in the Readable tool. It can be seen in Table 4.4. Additionally, the

averages of each vertex are identified in both RNN and LSTM (I RNN = 28.43 %, I LSTM = 77.71 %, II RNN = 16.15 %, II LSTM = 70.46 %, III RNN = 24.8 % and III LSTM = 77.05 %) respectively. We obtain the results of the metrics applied to the text generated by the RNN - LSTM algorithms. We can identify these sentences'consistency versus the original data representing the baseline against which those results will be compared to identify the most appropriate technique to build fraud-related phrases.



**Table 4.4:** Results in percentages of analysis using Readeable tool on data generated by RNN and LSTM algorithms

| Cantidad | I | | II | | III | |
|---|---|---|---|---|---|---|
| | RNN | LSTM | RNN | LSTM | RNN | LSTM |
| 100 | 18.7 % | 76.7 % | 14.3 % | 71 % | 23.3 % | 77.7 % |
| 200 | 31.5 % | 78.1 % | 19.5 % | 70.8 % | 25.6 % | 77.2 % |
| 300 | 28.3 % | 77.6 % | 17.9 % | 69.8 % | 20.9 % | 77.1 % |
| 400 | 31.6 % | 77.7 % | 16.7 % | 70 % | 21.7 % | 77.4 % |
| 500 | 29.1 % | 77.4 % | 16.4 % | 70.3 % | 23.9 % | 76.8 % |
| 600 | 31.2 % | 77.8 % | 16.2 % | 70.8 % | 23.8 % | 76.7 % |
| 700 | 31.4 % | 77.8 % | 15.1 % | 70.8 % | 25.2 % | 76.8 % |
| 800 | 30.7 % | 77.8 % | 15.6 % | 70.4 % | 25.9 % | 76.8 % |
| 900 | 27 % | 77.8 % | 15.9 % | 70.4 % | 25.7 % | 76.9 % |
| 1000 | 27 % | 78.1 % | 15.4 % | 70.4 % | 26.7 % | 77 % |
| 1500 | 26.2 % | 78 % | 14.6 % | 70.4 % | 25.5 % | 77.2 % |
| Avg. | 28.43 % | 77.71 % | 16.15 % | 70.46 % | 24.8 % | 77.05 % |

Fig. 4.7 shows the comparison of the obtained metrics; the scores related to the original data corresponding to pressure, opportunity, and rationalization are represented by orange graphs. For the data generated by the RNN algorithm in the three mentioned vertexes, they are yellow graphs. Furthermore, finally, the data generated by the LSTM algorithm by blue graphics. First, we can see the comparison between the results obtained when analyzing each of the vertexes of the fraud triangle for the RNN Algorithm and the original dataset; the data collected by the algorithm have a score below 40 %, and the original dataset a score above 70 %, which shows that specifically, the RNN algorithm is inefficient. Regarding the scores obtained by the LSTM Algorithm for each vertex of the fraud triangle, we can see that the data collected by the algorithm has a score above 70 %, like the whole of the original data. Therefore, the text generated by the LSTM algorithm has consistency. Let's leave aside the score for a moment and focus on the number of sentences analyzed. We can affirm that a deep learning algorithm provides better results when working with large amounts of data. If we observe Fig. 4.7, we can see that from the 1000 data, the scores stabilize and do not make significant variations, while before the 1000 data, these scores are highly variable.

The average of the scores obtained by the different data sources (Original, RNN, and LSTM)

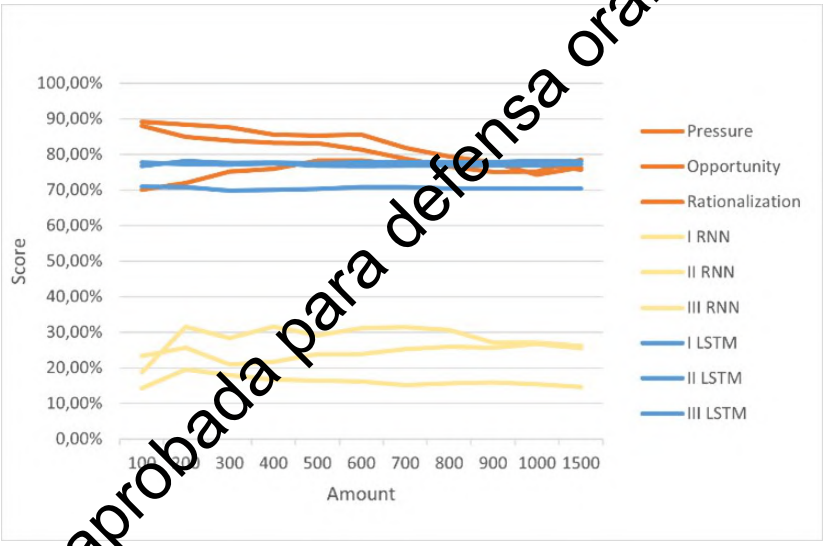


**Figure 4.7:** Score comparison between the data generated by RNN and LSTM algorithms against the original data.

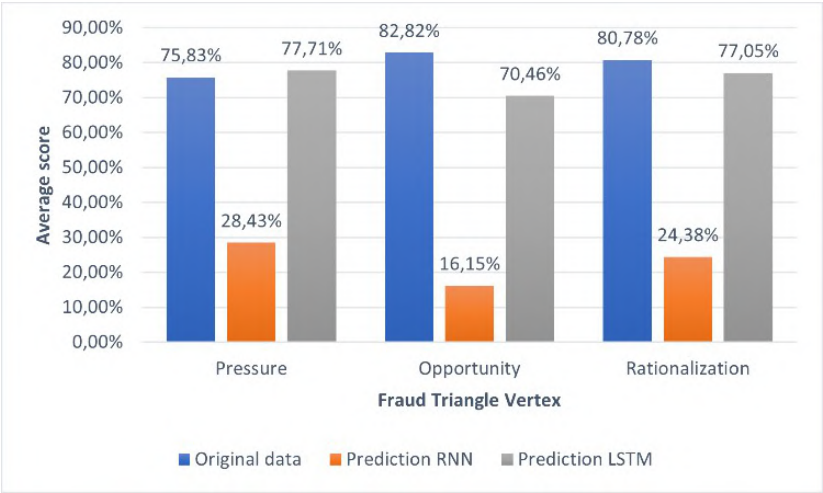


**Figure 4.8:** Comparison of averages obtained by the algorithms (RNN-LSTM) and the original data at the vertices of the fraud triangle.

at the fraud triangle's vertices indicate that the data generated by the RNN algorithm (orange color) is inefficient with percentages that do not exceed 28.43 % in the best case. The opposite occurs with the LSTM algorithm (gray color) with satisfactory results and even surpasses the original data (blue color) with 77.71 % for Pressure. In comparison, Opportunity and Rationalization are above 70 %, as can be seen in Fig. 4.8.

### 4.4.6 Implementation Recommendations

Under constrained resources, such as the hosts used in the development and testing phase, the Google Colab platform provides resources that allow storing and processing of large

amounts of data used in this research. The number of interactions to achieve optimal results varies between the different proposed algorithms; however, the model results'difference without oversampling is significant. For this reason, it is best to perform oversampling before training the algorithm for a small dataset or with unbalanced data. Another case to consider is not to overtrain the Neural Network since a more significant number of iterations does not always mean greater precision in the results. For example, in the LSTM algorithm developed in this project, when trying 150 iterations, the results did not differ much from 100. The variety of optimizers available for the implementation of neural networks can make developers hesitate between them for their purposes. We corroborate that the Adam algorithm (Adaptive moment estimation) works well for text generation. Adam combines the benefits of the AdaGrad and RMSProp algorithms.



## 4.5 CONCLUSION

The limited availability of datasets for the analysis and study of fraud presents a challenge in developing tools for its detection. This paper has presented a methodology for the generation of uniformly distributed synthetic data based on the fraud triangle theory. We used RNN and LSTM, an original dataset, and a data dictionary built on the fraud triangle's vertices (pressure, opportunity, and rationalization) to generate fraud-related sentences.

We compared the consistency of original and synthetically generated data distributions based on their readability and grammar. Initially, the original data's consistency was analyzed, obtaining a score higher than 70 % as a baseline. Our results show that the original data's consistency has a score higher than 70 % as a baseline, which serves as the baseline. The synthetic dataset generated with the RNN algorithm is deficient and has a consistency below 40 %. On the contrary, the LSTM algorithm maintains a consistency level higher than 70 % and is similar to the original data's score. As future work, we will propose a fraud predictor system that employs machine learning algorithms.

# REFERENCES

[1] Acfe-spain.com. Acfe association of certified fraud examiners capítulo españa. Available: https://acfe-spain.com. Accedido 06-11-2019.

[2] D. Cressey. *Other people's money. Montclair, NJ: Patterson Smith*. 1973.

[3] Jiaqi Guan, Runzhe Li, Sheng Yu, and Xuegong Zhang. A method for generating synthetic electronic medical record text. *IEEE/ACM transactions on computational biology and bioinformatics*, 18(1):173–182, 2019.

[4] Richard A DeMillo, A Jefferson Offutt, et al. Constraint-based automatic test data generation. *IEEE Transactions on Software Engineering*, 17(9):900–910, 1991.

[5] Mukesh Mann, Om Praksah Sangwan, Pradeep Tomar, and Shivani Singh. Automatic goal-oriented test data generation using a genetic algorithm and simulated annealing. In *2016 6th International Conference-Cloud System and Big Data Engineering (Confluence)*, pages 83–87. IEEE, 2016.

[6] Shweta Rani and Bharti Suri. An approach for test data generation based on genetic algorithm and delete mutation operators. In *2015 Second International Conference on Advances in Computing and Communication Engineering*, pages 714–718. IEEE, 2015.

[7] Georgia Albuquerque, Thomas Lowe, and Marcus Magnor. Synthetic generation of high-dimensional datasets. *IEEE transactions on visualization and computer graphics*, 17(12):2317–2324, 2011.

[8] Bing Wang, Puripant Ruchikachorn, and Klaus Mueller. Sketchpadn-d: Wydiwyg sculpting and editing in high-dimensional space. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2060–2069, 2013.

[9] Bum Chul Kwon, Hannah Kim, Emily Wall, Jaegul Choo, Haesun Park, and Alex Endert. Axisketcher: Interactive nonlinear axis mapping of visualizations through user drawings. *IEEE transactions on visualization and computer graphics*, 23(1):221–230, 2016.

[10] Runzong Liu, Bin Fang, Yuan Yan Tang, and Patrick PK Chan. Synthetic data generator for classification rules learning. In *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*, pages 357–361. IEEE, 2016.

[11] Pengyue J Lin, Behrokh Samadi, Alan Cipolone, Daniel R Jeske, Sean Cox, Carlos Rendón, Douglas Holt, and Rui Xiao. Development of a synthetic data set generator for building and testing information discovery systems. In *Third International Conference on Information Technology: New Generations (ITNG'06)*, pages 707–712. IEEE, 2006.

[12] Daniel R Jeske, Pengyue J Lin, Carlos Rendon, Rui Xiao, and Behrokh Samadi. Synthetic data generation capabilties for testing data mining tools. In *MILCOM 2006-2006 IEEE Military Communications conference*, pages 1–6. IEEE, 2006.

[13] Marden Pasinato, Carlos Eduardo Mello, Marie-Aude Aufaure, and Geraldo Zimbrao. Generating synthetic data for context-aware recommender systems. In *2013 BRICS Congress on Computational Intelligence and 11th Brazilian Congress on Computational Intelligence*, pages 563–567. IEEE, 2013.

[14] D García and M Millán. A prototype of synthetic data generator. In *2011 6th Colombian Computing Congress (CCC)*, pages 1–6. IEEE, 2011.

[15] Felix Brodkorb, Manuel Kopp, Arjan Kuijper, and Tatiana Von Landesberger. A modular rule-based visual interactive creation of tree-shaped geo-located networks. In *2016 12th International Conference on Signal-image Technology & Internet-based Systems (sitis)*, pages 397–403. IEEE, 2016.

[16] Xiaowei Ying and Xintao Wu. Graph generation with prescribed feature constraints. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pages 966–977. SIAM, 2009.

[17] Can Yang, Sixuan Ren, Yong Liu, Houwei Cao, Qihu Yuan, and Guoqiang Han. Personalized channel recommendation deep learning from a switch sequence. *IEEE Access*, 6:50824–50838, 2018.

[18] Brian C Hosler, Xinwei Zhao, Owen Mayer, Chen Chen, James A Shackleford, and Matthew C Stamm. The video authentication and camera identification database: A new database for video forensics. *IEEE Access*, 7:76937–76948, 2019.

[19] Hangxia Zhou, Yujin Zhang, Lingfan Yang, Qian Liu, Ke Yan, and Yang Du. Short-term photovoltaic power forecasting based on long short term memory neural network and attention mechanism. *Ieee Access*, 7:78063–78074, 2019.

[20] Hamdi Altaheri, Mansour Alsulaiman, and Ghulam Muhammad. Date fruit classification for robotic harvesting in a natural environment using deep learning. *IEEE Access*, 7:117115–117133, 2019.

[21] Ahmadreza Argha, Ji Wu, Steven W Su, and Branko G Celler. Blood pressure estimation from beat-by-beat time-domain features of oscillometric waveforms using deep-neural-network classification models. *IEEE Access*, 7:113427–113439, 2019.

[22] Xun Zhu, Chen Lyu, and Donghong Ji. Keyphrase generation with copynet and semantic web. *IEEE Access*, 8:44202–44210, 2020.

[23] Noorhayati Mansor and Rabiu Abdullahi. Fraud triangle theory and fraud diamond theory. understanding the convergent and divergent for future research. *International Journal of Academic Research in Accounting, Finance and Management Science*, 1(4):38–45, 2015.

[24] Shaio Yan Huang, Chi-Chen Lin, An-An Chiu, and David C Yen. Fraud detection using fraud triangle risk factors. *Information Systems Frontiers*, 19(6):1343–1356, 2017.

[25] Trivellore E Raghunathan, Jerome P Reiter, and Donald B Rubin. Multiple imputation for statistical disclosure limitation. *Journal of official statistics*, 19(1):1, 2003.

[26] Jennifer Taub, Mark Elliot, and Joseph W Sakshaug. The impact of synthetic data generation on data utility with application to the 1991 uk samples of anonymised records. *Transactions on Data Privacy*, 13(1):1–23, 2020.

[27] Alicia R. Fernández. “los datos sintéticos, la clave para mejorar la inteligencia artificial”, addison wesley college, 1997. Available: https://www.ticbeat.com/tecnologias/los-datos-sinteticos-la-clave-para-mejorar-la-inteligencia-artificial/. Accedido 07-07-2020.

[28] SB Kotsiantis and PE Pintelas. Mixture of expert agents for handling imbalanced data sets. *Annals of Mathematics, Computing & Teleinformatics*, 1(1):46–55, 2003.

[29] Mufti Mahmud, Mohammed Shamim Kaiser, Amir Hussain, and Stefano Vassanelli. Applications of deep learning and reinforcement learning to biological data. *IEEE transactions on neural networks and learning systems*, 29(6):2063–2079, 2018.

[30] V Golovko, A Kroshchanka, and D Treadwell. The nature of unsupervised learning in deep neural networks: A new understanding and novel approach. *Optical memory and neural networks*, 25(3):127–141, 2016.

[31] Damián Jorge Matich. Redes neuronales: Conceptos básicos y aplicaciones. *Universidad Tecnológica Nacional, México*, 41:12–16, 2001.

[32] Ruiqin Bai, Junjun Zhao, Dengao Li, Xiaoyu Lv, Qiang Wang, and Biaokai Zhu. Rnn-based demand awareness in smart library using crfid. *China Communications*, 17(5):284–294, 2020.

[33] Daniele Ravì, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang-Zhong Yang. Deep learning for health informatics. *IEEE journal of biomedical and health informatics*, 21(1):4–21, 2016.

[34] S Dhananjay Kumar and DP Subha. Prediction of depression from eeg signal using long short term memory (lstm). In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1248–1253. IEEE, 2019.

[35] Jianjing Zhang, Peng Wang, Ruqiang Yan, and Robert X Gao. Long short-term memory for machine remaining life prediction. *Journal of manufacturing systems*, 48:78–86, 2018.

[36] Emilie Lundin, Håkan Kvarnström, and Erland Jonsson. A synthetic fraud data generation methodology. In *International Conference on Information and Communications Security*, pages 265–277. Springer, 2002.

[37] AuditNet. “using key word analysis of an organization’s big data for error and fraud detection”. urlhttps://www.auditnet.org/key-word-analytics.

# 5 COMPARATIVE ANALYSIS OF THE PERFORMANCE OF MACHINE LEARNING TECHNIQUES APPLIED TO REAL AND SYNTHETIC FRAUD-ORIENTED DATASETS

Marco Sánchez[1*], Luis Urquiza[2]

[1]Departamento de Informática y Ciencias de la Computación, Escuela Politécnica Nacional, Ladrón de Guevara E11-253, Quito 170517, Ecuador

[2]Departamento de Electrónica, Telecomunicaciones y Redes de Información, Escuela Politécnica Nacional, Ladrón de Guevara E11-253, Quito 170517, Ecuador

## 5.1 ABSTRACT

One of the most critical resources today is information, an intangible asset that has become a vital research source. On many occasions, access to data becomes a complex and challenging task. For many organizations, sharing information is often a risk in terms of security and privacy, especially if the data is sensitive. In response to this problem, synthetic data emerges as a valid alternative, generated by different methods and techniques from an original or real dataset, allowing the sharing of information very close to reality. In this work, an experiment is carried out that allows validating the efficiency of synthetic versus real datasets by applying a model that predicts possible fraud cases in a dataset based on machine learning algorithms LDA and Random Forest or Gradient Boosting. We compared the prediction performance of our model over the real and synthetic datasets using metric ROC-AUC curves. Our results show a similar behavior among the datasets in our model, suggesting a promising path in the use of synthetic datasets for this kind of applications.

**KEY WORDS:** Fraud, Real and Synthetic dataset, Classification Methods, AUC-ROC, Topic Modeling.

## 5.2 INTRODUCTION



Fraud is a global concern that affects both public and private institutions, and it encompasses a wide range of illegal actions, including deliberate deceit or misrepresentation. The Association of Certified Fraud Examiners (ACFE) defines fraud as “any purposeful or deliberate act of depriving someone of property or money by cunning, deception, or other unfair acts.” [1].

According to a Price Waterhouse Coopers investigation, 30 % of the organizations examined have already been victims of fraud. Furthermore, 80 % of their fraud was done within the company’s ranks, particularly in administrative departments such as accounting, operations, sales, and management, not to mention customer service relationships [2]. Often unknown within a corporation, fraud-related practices define a sequence of anomalies and illegal acts defined by fraudsters’purposeful deceit. Most discovered abnormalities result from a lack of internal control systems, and in such cases, fraudsters perpetrate fraud by leveraging the flaws [3]. Because humans commit fraud, it is closely related to their behavior. Therefore, understanding the motivations of perpetrators or their psychological and personality traits that lead them to cross ethical boundaries can provide a new perspective for fraud detection [4]. There is agreement that prevention should be a primary approach to reducing fraud through effective risk management. Avoiding fraud saves time and money since detecting it after it has occurred makes it almost impossible to recover what was stolen. To increase fraud prevention, companies must identify those factors that drive people to commit fraud and understand this behavior [5]. Numerous theories have tried to explain this issue, being Cressey’s Fraud Triangle Theory (FTT) and Wolf and Hermanson’s Fraud Diamond Theory (FDT) [23], the most referenced in this field. Both techniques examine in-depth the invoices that motivate committing fraud.

One of the most difficult challenges to the investigation and study of fraud is the lack of access to data linked to this issue. Except for studies conducted by private entities such as the Federal Bureau of Investigation (FBI) and ACFE, information with evidence of fraudulent activities associated with fraud theory, in which communications related to pressure, opportunity, and rationalization are observed, is incipient in the scientific community. They were successful in obtaining data related to this topic of research. For the development of fraud prevention methods, fraud-related data is essential. Actual datasets are scarce due to infringements of copyright and intellectual property. Due to the difficulties of acquiring this

sensitive information, the fabrication of synthetic data is a viable approach for acquiring fraud data. According to several experts, synthetic data makes ML and AI quicker and their algorithms more effective at predicting fraudulent behavior, particularly when acquiring actual data is costly or difficult [6].

The scientific community often employs synthetic data production. These data are often created to fit particular criteria not present in the original data. Researchers may manipulate data more freely and test broader settings and scenarios in their applications by creating synthetic datasets [7]. In experimental investigations, synthetic datasets that follow statistical distributions and data from real-world applications are used as test datasets. Synthetic datasets allow testing an algorithm's or data structure's behavior under specific conditions or in extreme situations. Also, for testing scalability, synthetic datasets are often suitable [8].

This article analyzes the validity of synthetic data generated through neural networks and tools available on the internet, which synthesize data based directly on real data of interest. The real data was obtained through simulation with students from the Escuela Politécnica Nacional (EPN). Validation of the use of synthetic data for research requires a comparison of results derived from synthetic data with those based on original data.

Through a model that allows the detection of suspected fraud behaviors, which uses a theory to analyze this phenomenon from the point of view of human behavior known as FTT, plus modeling of topics and automatic learning algorithms as classification methods allow alerting on the possibility of fraud in a dataset. This model will be used to carry out a comparative study of real and synthetic datasets. In this work, the validation of three datasets generated by different methods is carried out using the mentioned model, in which topic modeling is applied, which is a widely used approach in text mining and provides a complete representation of a corpus through the inference of latent content variables called topics. This technique assigns a probability to a text or document belonging to a specific topic [9]. Different classification methods will use the probability that a document belongs to a topic to identify which technique is more compatible with topic modeling and efficiently identify phrases suspected of fraud. The AUC-ROC curve was used to measure the classification models'performance. As a result, it was observed that the Random Forest (RF) and Gradient Boosting algorithms were the most efficient in predicting possible fraud cases, and these methods will be used to compare the datasets under study.

The rest of this paper is organized as follows: Section 5.3 presents a literature review in the

area of dataset comparison. Section 5.4 describes the data preparation and the methodology used in this work. Next, Section 5.5 presents the experiment and the results. Finally, Section 5.6 presents the conclusions and future work.

## 5.3 RELATED WORK

Many areas of study use synthetically generated data, from data mining to software engineering to artificial intelligence. However, few works are in charge of comparative analysis of synthetic datasets against real datasets based on their performance applying classification methods. In this sense, the following studies were found in the literature contributing to this topic of study.

In [10], signal detection performance based on synthetic training data is compared with the performance of real-world training images. With synthetic and real data and a configurable number of training samples, Viola-Jones detectors are constructed for four distinct traffic lights. We test and evaluate detectors. The goal of [11] is to investigate whether synthetic data can be used as a reliable substitute for real-world data in machine learning systems. This research evaluates the performance of synthetic datasets when used to train machine learning models. Using three object identification methods, [12] verified the synthetic data for model pretraining and data augmentation to examine the synthetic dataset's utility. Our findings demonstrate that the synthetic dataset considerably enhances model pretraining and data augmentation for small and medium-sized real-world datasets, illustrating the utility and promise of synthetic data in aerial imagery. In [13], they validate five studies on the omission of suggested medicine, the influence of time to procedure, and hospitalization measures on survival after discharge, imaging risks, and diabetes therapies. Institutional review board (IRB) approval was acquired to utilize real data, allowing real and synthetic data comparison. These studies evaluated the accuracy and precision of synthetic patient data-based estimations. On the other hand, [11] experimented with studying the validity of performing machine learning on synthetic data. They compared evaluation metrics from machine learning models trained on synthetic data with metrics from machine learning models trained on the corresponding real data by generating a fully synthetic dataset through subsampling a synthetically generated population and generating a partially synthetic dataset by obtaining the values of sensitive attributes.

The authors of [14] studied these techniques using different dataset synthesizers such as

linear regression, decision tree, random forest, and neural network. They evaluated the effectiveness of these techniques towards the amounts of utility they preserve and the disclosure risks they suffer. The features of the synthetic data are compared to those of the original data in the work proposed by [15], and a model demonstrating how the synthetic data may be utilized to create and improve a standard learning analysis is shown. [7], a method for producing synthetic microdata utilizing the publicly accessible tool Benerator to introduce a new domain for data generation based on census-based personal information is discussed. In addition, they examine the distributions of the original and synthetic data, revealing that the synthetic dataset maintains a high degree of accuracy in contrast to the original distribution. In this work [16], the authors analyze a cancer clinical trial to show how synthetic data may be used to get the same conclusions from real data. These findings imply that synthetic data may act as a stand-in for real data, increasing the accessibility of relevant clinical trial data to researchers. Unlike previous research, our work will compare and evaluate the performance of different synthetic datasets to identify if they can be a reliable replacement for actual data by using a tool that detects possible fraud cases. This model identifies suspicious fraud behaviors in a dataset through topic modeling techniques and classification methods, which, aligned with the FTT, allow addressing this phenomenon from a sociological point of view, associating the different behaviors found to the vertices of this theory.



## 5.4 METHODOLOGY

### 5.4.1 Dataset Selection

Finding evidence confirming the occurrence of fraud becomes challenging when studying and analyzing this phenomenon. Whether due to its importance or sensitivity, the corporations and organizations that own this source of information protect it. Often due to their confidentiality rules, which restrict access to this resource. Researchers typically use real data for analysis and experimentation in their research. However, synthetically generated datasets can solve this problem when access to this information is limited or non-existent [61, 17]. For this work, two synthetic datasets were generated from a real fraud-oriented dataset created at the EPN. This was done through a controlled experiment with EPN students, for which a data dictionary called “Textual Survey Word List 103115” was used, acquired from the company Audinet [18], which contains words related to the three vertices of the FTT, “Pressure,

Opportunity and Rationalization," Which was used to create this initial or real dataset containing phrases related to fraud. This real dataset, named for this paper "Students", comprises 14,226 records balanced in two classes: fraud and non-fraud (7113 × 7113). Each sentence belongs to one of two classes: fraud, represented by a 1, or not fraud, represented by a 0. This initial dataset served as a seed to feed a neural network and tools available on the internet to generate two synthetic datasets, which were used to feed the model to predict possible cases of fraud mentioned above.



### 5.4.2 Generating Synthetic Data

To analyze any phenomenon that needs to be studied, it is recommended to have real data. However, without this resource, the data generated synthetically by some simulation tool becomes a valid alternative. The generation of synthetic data is a complex task. It demands resources for its execution, so it is necessary to use an adequate methodology that optimizes this work and establishes an adequate procedure that allows the execution of the related tasks. The methodology proposed by Lundin et al. [19] was taken as a reference for the generation of synthetic datasets, adapting it to the required needs and depending on the tools used. Different strategies were used to generate the synthetic datasets. As requirements, the characteristics of the real dataset were established as functional parameters, referring to the number of records and classes used. The first synthetic dataset, named "WebScraping", was constructed from the use of various keywords related and unrelated to the FTT, in the same proportion as the real dataset (7113 x 7113) for fraud and non-fraud, respectively, using the phrases related to fraud; the dictionary "Textual Survey Word List 103115" and for phrases not related to fraud words not related to this phenomenon. Using different online tools to generate text, like [20, 21, 22]; Phrases were obtained that included the selected keywords. These tools allow sentences to be generated from a specific word with a well-defined grammatical and semantic structure. Finally, a web scraping tool, "Firefox Addon," allows us to save the generated results and export them in CSV format for processing. The process followed to generate this dataset is shown in Fig. 5.1.

For the second synthetic dataset, named "Neural-Networks," the methodology established by [23] was used, in which a portion of sentences of the real or initial dataset "Students," was used as input for generating text related and not related to fraud, which, as in the previous synthetic dataset "WebScraping" kept the same parameters in which the real dataset

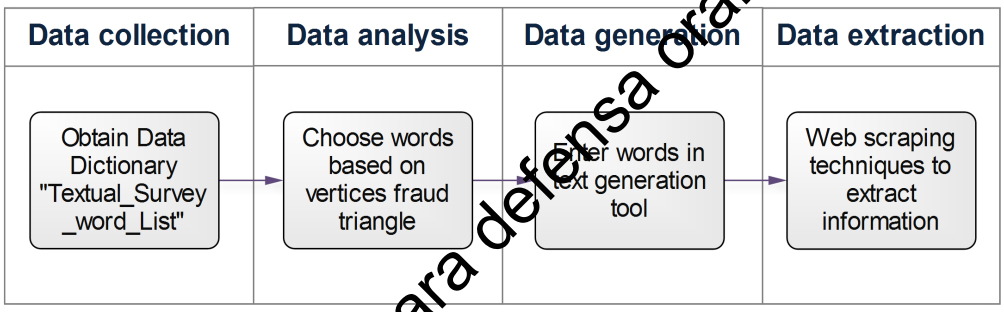


**Figure 5.1:** Flow chart used to generate the synthetic dataset named “WebScraping”.



was built. The next step is to review the data collected using [19] exploratory data analysis (EDA). In addition, essential characteristics are identified and valuable parameters for fraud detection. Next, relevant parameters are identified in the input data, and one way to identify these parameters is to study the characteristics necessary for fraud detection. These features must have properties related to the FTT. The result of this stage will allow the identification of a suitable profile to analyze fraudulent activities. Finally, the real dataset will be downsampled to balance the minority class with the majority class. The initial dataset, composed of phrases identified as fraud and non-fraud, will be the input for the text generation algorithms by applying deep learning algorithms such as recurrent neural networks (RNN) and long short-term memory (LSTM). The process followed to generate this dataset is shown in Fig. 5.2.

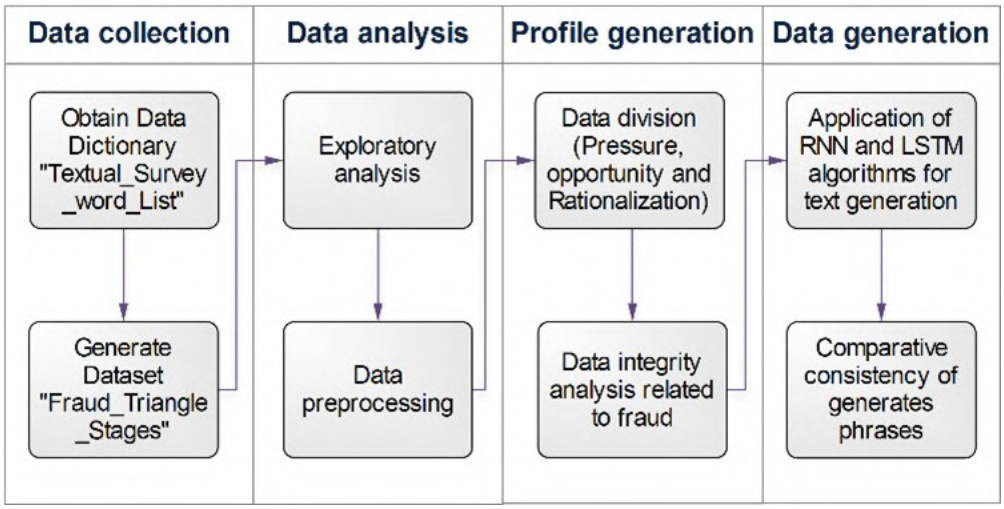


**Figure 5.2:** Flow chart used to generate the synthetic dataset named “Neural-Networks”.

**Table 5.1:** Probabilities per topic obtained by LDA of the study datasets (Students, WebScraping and Neural-Networks).

| Docs | Students | | | | DT | WebScraping | | | | DT | Neural-Networks | | | | DT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | | 1 | 2 | 3 | 4 | | 1 | 2 | 3 | 4 | |
| 0 | 0.08 | 0.08 | 0.75 | 0.08 | 2 | 0.91 | 0.03 | 0.03 | 0.03 | 0 | 0.36 | 0.03 | 0.03 | 0.58 | 3 |
| 1 | 0.62 | 0.13 | 0.13 | 0.13 | 0 | 0.83 | 0.06 | 0.05 | 0.05 | 0 | 0.6 | 0.31 | 0.04 | 0.05 | 0 |
| 2 | 0.05 | 0.05 | 0.85 | 0.05 | 2 | 0.02 | 0.66 | 0.02 | 0.3 | 1 | 0.93 | 0.02 | 0.02 | 0.02 | 0 |
| 3 | 0.05 | 0.05 | 0.85 | 0.05 | 2 | 0.89 | 0.04 | 0.04 | 0.04 | 0 | 0.56 | 0.04 | 0.37 | 0.04 | 0 |
| 4 | 0.06 | 0.06 | 0.56 | 0.31 | 2 | 0.95 | 0.02 | 0.02 | 0.02 | 0 | 0.02 | 0.02 | 0.74 | 0.23 | 2 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 14222 | 0.05 | 0.05 | 0.25 | 0.65 | 3 | 0.02 | 0.02 | 0.68 | 0.28 | 2 | 0.04 | 0.19 | 0.04 | 0.73 | 3 |
| 14223 | 0.08 | 0.08 | 0.42 | 0.42 | 2 | 0.2 | 0.04 | 0.26 | 0.49 | 3 | 0.4 | 0.07 | 0.07 | 0.47 | 3 |
| 14224 | 0.04 | 0.04 | 0.04 | 0.89, | 3 | 0.34 | 0.03 | 0.61 | 0.03 | 2 | 0.05 | 0.05 | 0.05 | 0.85 | 3 |
| 14225 | 0.06 | 0.06 | 0.06 | 0.81 | 3 | 0.75 | 0.02 | 0.2 | 0.02 | 0 | 0.25 | 0.05 | 0.05 | 0.65 | 3 |
| 14226 | 0.06, | 0.06 | 0.06 | 0.81 | 3 | 0.44 | 0.15 | 0.39 | 0.02 | 0 | 0.5 | 0.06 | 0.06 | 0.37 | 0 |

### 5.4.3 Topic modeling and Classification Methods used in Real and Synthetic datasets

Taking as reference the model proposed by [24], in which they propose identifying hidden patterns within a dataset that may be related to fraud. To achieve this, they develop a model to predict if a specific phrase belongs to one of these categories (pressure, opportunity, rationalization, and others). If it matches one of the first three, this phrase is suspected of fraud. To detect suspicious fraud-related patterns, in the first phase, they perform topic modeling (unsupervised learning) on an unstructured dataset [25]. They select Latent Direchlet Allocation (LDA) as the best topic model. Then, based on the resulting coherence value, which indicates the level of semantic similarity between words on a topic [26], they determine the appropriate number of topics or k value. This value is an input parameter needed to obtain a topic model in LDA. They determine a value of k = 4 in their study. Once the appropriate value of k is obtained, LDA is applied to the study corpus. We proceed to extract the probabilities that the documents belong to specific topic values provided by the algorithm that will be useful to feed classification methods and try to predict phrases related to fraud, as can be seen in Table 5.1.

In the second phase, with the probabilities that the documents belong to a specific topic (obtained from the LDA model) from the datasets, the records are labeled with 1 or 0 to indicate whether or not it is related to fraud. Documents grouped by dominant topic (DT) and their indicator related to fraud or no fraud are selected to build new datasets (T1, T2, T3, and T4), as can be seen in the Tables 5.2, 5.3 and 5.4, related with the different study datasets “Students, WebScraping and Neural-Networks”. This new representation of the datasets will be used as input for different classification algorithms, whose resulting prediction models will

**Table 5.2:** Segmentation of probabilities by Dominant Topic (DT) and labeling fraud=1 and no fraud=0 (Students Dataset).

| DT 1 | | | | F | DT 2 | | | | F | DT 3 | | | | F | DT 4 | | | | F |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | | 1 | 2 | 3 | 4 | | 1 | 2 | 3 | 4 | | 1 | 2 | 3 | 4 | |
| 0.62 | 0.13 | 0.13 | 0.13 | 1 | 0.05 | 0.62 | 0.28 | 0.05 | 1 | 0.08 | 0.08 | 0.75 | 0.08 | 1 | 0.06 | 0.38 | 0.06 | 0.49 | 1 |
| 0.5 | 0.17 | 0.29 | 0.04 | 1 | 0.05 | 0.65 | 0.25 | 0.05 | 1 | 0.05 | 0.05 | 0.85 | 0.05 | 1 | 0.04 | 0.05 | 0.21 | 0.71 | 1 |
| 0.44 | 0.29 | 0.24 | 0.03 | 1 | 0.04 | 0.45 | 0.34 | 0.18 | 1 | 0.05 | 0.05 | 0.85 | 0.05 | 1 | 0.13 | 0.13 | 0.13 | 0.62 | 1 |
| 0.6 | 0.31 | 0.04 | 0.04 | 1 | 0.44 | 0.45 | 0.06 | 0.06 | 1 | 0.06 | 0.06 | 0.56 | 0.31 | 1 | 0.21 | 0.23 | 0.04 | 0.52 | 1 |
| 0.62 | 0.13 | 0.13 | 0.13 | 1 | 0.35 | 0.53 | 0.06 | 0.06 | 1 | 0.06 | 0.06 | 0.81 | 0.06 | 1 | 0.06 | 0.06 | 0.31 | 0.56 | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 0.56 | 0.06 | 0.06 | 0.31 | 0 | 0.05 | 0.52 | 0.05 | 0.38 | 0 | 0.06 | 0.06 | 0.56 | 0.31 | 0 | 0.08 | 0.08 | 0.08 | 0.75 | 0 |
| 0.42 | 0.08 | 0.08 | 0.42 | 0 | 0.08 | 0.42 | 0.08 | 0.42 | 0 | 0.03 | 0.16 | 0.41 | 0.41 | 0 | 0.05 | 0.05 | 0.25 | 0.65 | 0 |
| 0.25 | 0.25 | 0.25 | 0.25 | 0 | 0.08 | 0.42 | 0.08 | 0.42 | 0 | 0.08 | 0.08 | 0.42 | 0.42 | 0 | 0.04 | 0.04 | 0.04 | 0.89 | 0 |
| 0.46 | 0.06 | 0.07 | 0.41 | 0 | 0.06 | 0.56 | 0.06 | 0.31 | 0 | 0.06 | 0.29 | 0.33 | 0.31 | 0 | 0.06 | 0.06 | 0.06 | 0.81 | 0 |
| 0.42 | 0.08 | 0.08 | 0.42 | 0 | 0.08 | 0.42 | 0.08 | 0.42 | 0 | 0.08 | 0.08 | 0.42 | 0.42 | 0 | 0.06 | 0.06 | 0.06 | 0.81 | 0 |

**Table 5.3:** Segmentation of probabilities by Dominant Topic (DT) and labeling fraud=1 and no fraud=0 (WebScraping Dataset).

| DT 1 | | | | F | DT 2 | | | | F | DT 3 | | | | F | DT 4 | | | | F |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | | 1 | 2 | 3 | 4 | | 1 | 2 | 3 | 4 | | 1 | 2 | 3 | 4 | |
| 0.91 | 0.03 | 0.03 | 0.03 | 1 | 0.02 | 0.66 | 0.02 | 0.03 | 1 | 0.41 | 0.05 | 0.49 | 0.05 | 1 | 0.07 | 0.06 | 0.06 | 0.81 | 1 |
| 0.84 | 0.06 | 0.05 | 0.05 | 1 | 0.07 | 0.8 | 0.06 | 0.06 | 1 | 0.02 | 0.21 | 0.74 | 0.02 | 1 | 0.41 | 0.01 | 0.01 | 0.57 | 1 |
| 0.89 | 0.04 | 0.04 | 0.04 | 1 | 0.37 | 0.46 | 0.03 | 0.14 | 1 | 0.05 | 0.44 | 0.45 | 0.05 | 1 | 0.02 | 0.02 | 0.47 | 0.49 | 1 |
| 0.95 | 0.02 | 0.02 | 0.02 | 1 | 0.04 | 0.88 | 0.04 | 0.04 | 1 | 0.38 | 0.18 | 0.42 | 0.02 | 1 | 0.13 | 0.13 | 0.13 | 0.62 | 1 |
| 0.87 | 0.04 | 0.04 | 0.04 | 1 | 0.04 | 0.9 | 0.03 | 0.03 | 1 | 0.32 | 0.19 | 0.45 | 0.04 | 1 | 0.33 | 0.19 | 0.01 | 0.46 | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 0.84 | 0.05 | 0.06 | 0.05 | 0 | 0.04 | 0.58 | 0.04 | 0.34 | 0 | 0.04 | 0.04 | 0.69 | 0.22 | 0 | 0.04 | 0.21 | 0.04 | 0.71 | 0 |
| 0.92 | 0.03 | 0.03 | 0.03 | 0 | 0.03 | 0.64 | 0.32 | 0.02 | 0 | 0.08 | 0.08 | 0.74 | 0.1 | 0 | 0.08 | 0.08 | 0.09 | 0.75 | 0 |
| 0.89 | 0.04 | 0.04 | 0.04 | 0 | 0.05 | 0.84 | 0.05 | 0.05 | 0 | 0.34 | 0.05 | 0.56 | 0.05 | 0 | 0.05 | 0.25 | 0.05 | 0.65 | 0 |
| 0.75 | 0.02 | 0.02 | 0.02 | 0 | 0.04 | 0.88 | 0.04 | 0.04 | 0 | 0.02 | 0.02 | 0.68 | 0.28 | 0 | 0.05 | 0.05 | 0.05 | 0.85 | 0 |
| 0.44 | 0.15 | 0.39 | 0.02 | 0 | 0.02 | 0.94 | 0.02 | 0.02 | 0 | 0.34 | 0.03 | 0.61 | 0.03 | 0 | 0.2 | 0.04 | 0.26 | 0.49 | 0 |



be used later to measure their performance and compare them. To compare the classifiers, it is essential to choose a good metric; they selected the area under the curve (AUC) since it is trendy when it is necessary to classify predictions and not necessarily obtain well-defined probabilities [27]. Random Forest (RF) and Gradient Boosting (GB) were the most efficient classification methods.

## 5.5 RESULTS

In the comparison of the classifiers, if the classes are balanced and there is no certainty that the classifier has chosen the best decision threshold, it is better to work with the AUC metric, which is equivalent to the probability that the classifier assigns the highest score to relevant classes compared to irrelevant ones [28]. The receiver operating characteristic (ROC) is a curve representing the rate of true positives against the rate of false positives, where the area determines the model's performance under the curve. The closer the AUC score is to 1, the better the model will distinguish between classes. In this work, the ROC curve was

**Table 5.4:** Segmentation of probabilities by Dominant Topic (DT) and labeling fraud=1 and no fraud=0 (Neural-Networks Dataset).

| DT 1 | | | | F | DT 2 | | | | F | DT 3 | | | | F | DT 4 | | | | F |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | | 1 | 2 | 3 | 4 | | 1 | 2 | 3 | 4 | | 1 | 2 | 3 | 4 | |
| 0.91 | 0.03 | 0.03 | 0.03 | 1 | 0.02 | 0.66 | 0.02 | 0.3 | 1 | 0.02 | 0.02 | 0.74 | 0.23 | 1 | 0.6 | 0.03 | 0.03 | 0.58 | 1 |
| 0.83 | 0.06 | 0.05 | 0.05 | 1 | 0.07 | 0.8 | 0.06 | 0.06 | 1 | 0.39 | 0.03 | 0.56 | 0.03 | 1 | 0.06 | 0.07 | 0.07 | 0.8 | 1 |
| 0.89 | 0.04 | 0.04 | 0.04 | 1 | 0.37 | 0.46 | 0.03 | 0.14 | 1 | 0.36 | 0.19 | 0.41 | 0.04 | 1 | 0.03 | 0.23 | 0.35 | 0.39 | 1 |
| 0.95 | 0.02 | 0.02 | 0.01 | 1 | 0.04 | 0.88 | 0.04 | 0.04 | 1 | 0.03 | 0.92 | 0.03 | 0.31 | 1 | 0.03 | 0.03 | 0.03 | 0.9 | 1 |
| 0.87 | 0.04 | 0.04 | 0.05 | 1 | 0.04 | 0.9 | 0.03 | 0.03 | 1 | 0.12 | 0.02 | 0.6 | 0.25 | 1 | 0.31 | 0.2 | 0.02 | 0.46 | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 0.84 | 0.05 | 0.06 | 0.05 | 0 | 0.04 | 0.58 | 0.04 | 0.34 | 0 | 0.03 | 0.03 | 0.54 | 0.39 | 0 | 0.08 | 0.36 | 0.08 | 0.47 | 0 |
| 0.92 | 0.03 | 0.03 | 0.02 | 0 | 0.03 | 0.64 | 0.12 | 0.02 | 0 | 0.03 | 0.03 | 0.63 | 0.31 | 0 | 0.04 | 0.19 | 0.04 | 0.73 | 0 |
| 0.88 | 0.04 | 0.04 | 0.04 | 0 | 0.05 | 0.84 | 0.05 | 0.05 | 0 | 0.08 | 0.08 | 0.42 | 0.42 | 0 | 0.4 | 0.07 | 0.07 | 0.47 | 0 |
| 0.75 | 0.02 | 0.2 | 0.02 | 0 | 0.04 | 0.87 | 0.04 | 0.04 | 0 | 0.31 | 0.06 | 0.32 | 0.31 | 0 | 0.05 | 0.05 | 0.05 | 0.85 | 0 |
| 0.44 | 0.15 | 0.39 | 0.02 | 0 | 0.02 | 0.95 | 0.02 | 0.02 | 0 | 0.05 | 0.05 | 0.47 | 0.42 | 0 | 0.25 | 0.05 | 0.05 | 0.65 | 0 |

**Table 5.5:** Performance measured with AUC, of RF and GB when classifying a document related or not to fraud within the study datasets (Students, WebScraping and Neuronal-Networks). T1, T2, T3, and T4 correspond to new datasets, each corresponding to a learned dominant topic of LDA.

| CM(AUC) | Students | | | | M | WebScraping | | | | M | Neural-Networks | | | | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | T1 | T2 | T3 | T4 | | T1 | T2 | T3 | T4 | | T1 | T2 | T3 | T4 | |
| Random Forest | 0.87 | 0.67 | 0.84 | 0.84 | **0.81** | 0.82 | 0.78 | 0.82 | 0.80 | **0.81** | 0.84 | 0.68 | 0.85 | 0.80 | **0.79** |
| Gradient Boosting | 0.87 | 0.68 | 0.86 | 0.84 | **0.81** | 0.85 | 0.79 | 0.83 | 0.83 | **0.83** | 0.85 | 0.70 | 0.92 | 0.82 | **0.82** |



used to represent the performance of different machine-learning models.

Once the model has been applied to the different study datasets, it can be seen that there is a similar behavior of the classifiers in the ROC-AUC curves by topic. In topics 0, 2, and 3, the performance values of the RF and GB algorithms obtained are very similar, with imperceptible differences. In contrast, in topic one, these differences are a little more visible, without this affecting the final average performance, as can be seen in Fig. 5.3.

In this context, about the real dataset “Students,” it was observed that the RF and the GB obtained an average AUC of 0.81 and 0.81, respectively. In contrast, the synthetic dataset generated by the internet “WebScraping” had a similar behavior when applying RF and GB, obtaining performance values with an average AUC of 0.81 and 0.83, respectively. Finally, in the second synthetic dataset, “Neural-Networks,” generated by deep learning, it can be seen that the RF and GB obtain an average AUC of 0.79 and 0.82, respectively, as can be seen in the Table 5.5.

These results suggest a similar behavior in the datasets analyzed based on the performance averages of the classifiers used, as can be seen in Fig. 5.4. Therefore, since synthetic datasets can be a very close alternative to the original data, it is feasible to produce a dataset that helps protect and protect information when it is confidential and difficult to access.

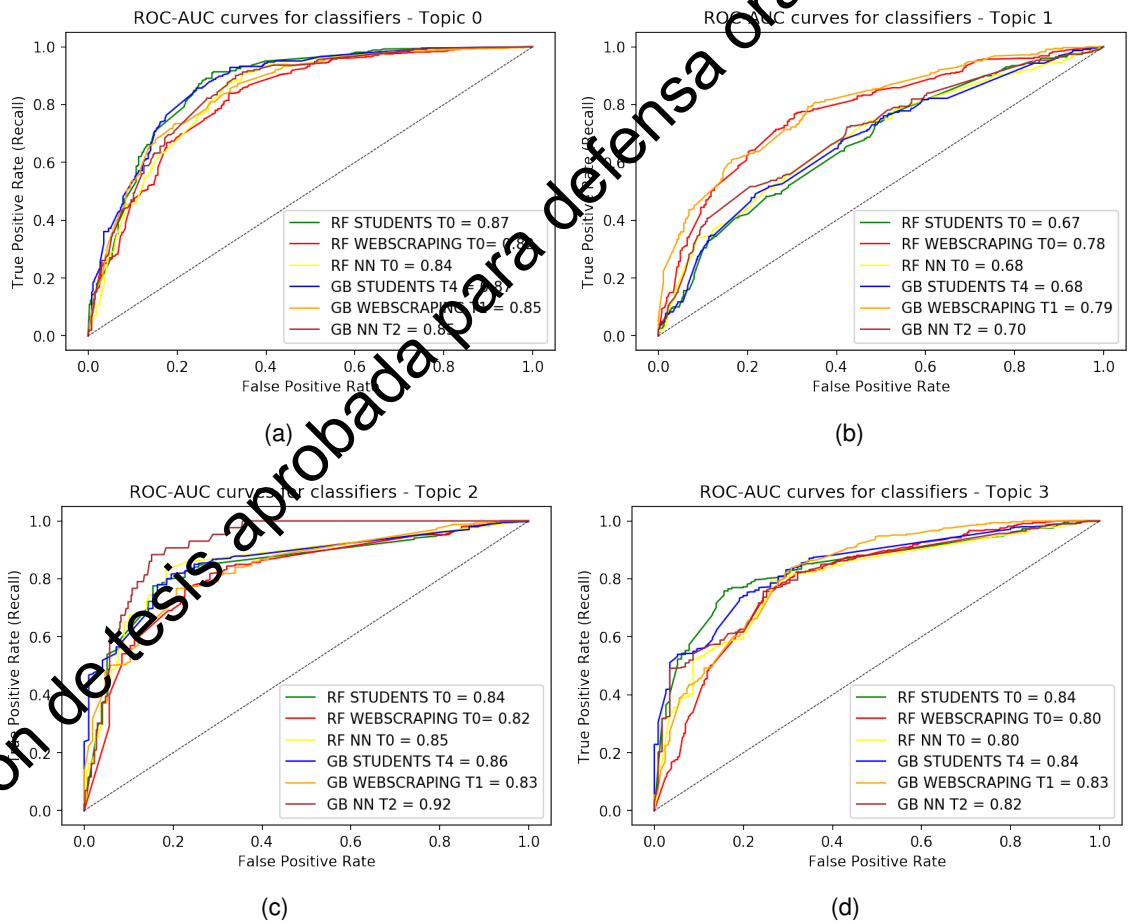




**Figure 5.3:** ROC curves of RF and GB classifiers for the real and synthetic datasets related to each dominant topic. (**a**) DT 1. (**b**) DT 2. (**c**) DT 3. (**a**) DT 4.

## 5.6 CONCLUSIONS

This work shows that the performance obtained by a detector of fraud-suspicious behavior based on machine learning algorithms used on the real dataset is similar to that obtained from synthetic datasets. These findings suggest that the results of models built using synthetic datasets may reflect behaviors obtained as if real data had been used. If more work supports this hypothesis, researchers can generate or use synthetic datasets with complete confidence that their results will have scientific validity. Synthetic datasets preserve the privacy and confidentiality of the information, allowing the development of predictive models to discover patterns without revealing confidential data, minimizing the risk of access to real data. It should also be mentioned that adequate evaluation metrics, which show the real behavior of the classifiers used, are essential since selecting the wrong one can be misleading in determining how the model behaves. In this case, according to the results obtained from the performance comparison, synthetic data is recommended to predict phrases sus-

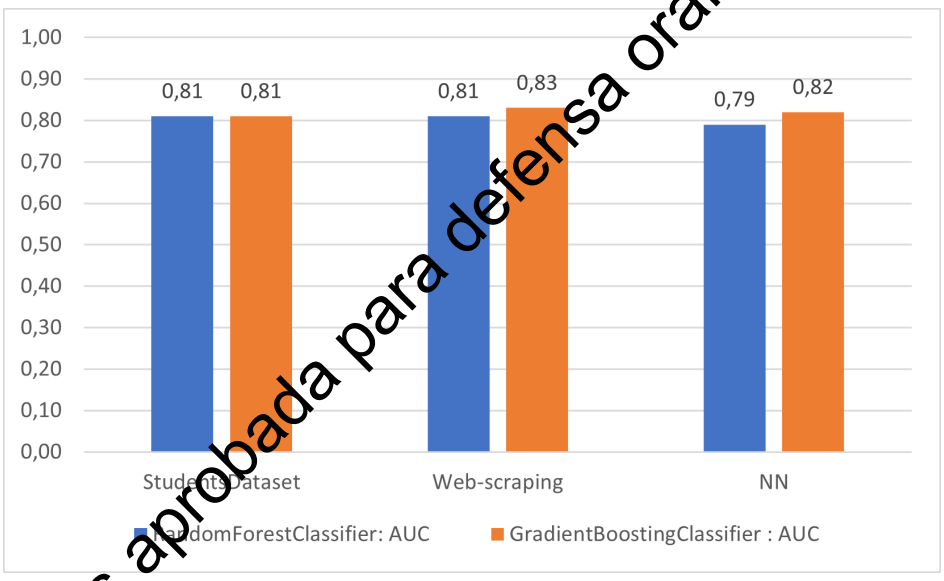


**Figure 5.4:** Best metrics obtained by the algorithms (Random Forest and Gradient Boosting) applied to the study datasets.



pected of fraud. As future work, it is proposed to conduct tests of the model for detecting fraud by applying deep learning algorithms and testing it with real and synthetic data to evaluate the performance and analyze if there is an improvement versus the classification methods.

## Acknowledgements

This work was sponsored by the Vicerrectorado de Investigación, Innovación y Vinculación from Escuela Politécnica Nacional. Marco Sánchez is the beneficiary of a teaching assistant fellowship from Escuela Politécnica Nacional for doctoral studies in Computer Science.

# REFERENCES


[1] Marco Sanchez, Jenny Torres, Patricio Zambrano, and Pamela Flores. FraudFind: Financial fraud detection by analyzing human behavior. In *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, jan 2018.

[2] PwC. (Date last accessed 15-July-2018).

[3] Prabin Kumar Panigrahi. A framework for discovering internal financial fraud using analytics. In *2011 International Conference on Communication Systems and Network Technologies*, pages 323–327. IEEE, 2011.

[4] G. Sayal, K.; Singh. hat role does human behaviour play in corporate frauds? In *Econ. Political Wkly.*, 2020.

[5] Thanasak Ruankaew. The fraud factors. 2013.

[6] Jiaqi Guan, Runzhe Li, Sheng Yu, and Xuegong Zhang. A method for generating synthetic electronic medical record text. *IEEE/ACM transactions on computational biology and bioinformatics*, 18(1):173–182, 2019.

[7] Vanessa Ayala-Rivera, Patrick McDonagh, Thomas Cerqueus, and Liam Murphy. Synthetic data generation using benerator tool. *arXiv preprint arXiv:1311.3312*, 2013.

[8] Thomas Brinkhoff. *Real and Synthetic Test Datasets*, pages 2339–2344. Springer US, Boston, MA, 2009.

[9] Pooja Kherwa and Poonam Bansal. Topic modeling: a comprehensive review. *EAI Endorsed transactions on scalable information systems*, 7(24), 2019.

[10] Andreas Møgelmose, Mohan M Trivedi, and Thomas B Moeslund. Learning to detect traffic signs: Comparative evaluation of synthetic and real-world datasets. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 3452–3455. IEEE, 2012.

[11] Rachel Heyburn, Raymond R Bond, Michaela Black, Maurice Mulvenna, Jonathan Wallace, Deborah Rankin, and Brian Cleland. Machine learning using synthetic and real data: similarity of evaluation metrics for different healthcare datasets and for different algorithms. In *Data Science and Knowledge Engineering for Sensing Decision Support: Proceedings of the 13th International FLINS Conference (FLINS 2018)*, pages 1281–1291. World Scientific, 2018.

[12] Boyong He, Xianjiang Li, Bo Huang, Enhui Gu, Weijie Guo, and Liaoni Wu. Unityship: A large-scale synthetic dataset for ship recognition in aerial images. *Remote Sensing*, 13(24):4999, 2021.

[13] Anat Reiner-Benaim, Ronit Almog, Yuri Gorelik, Irit Hochberg, Laila Nassar, Tanya Mashiach, Mogher Khamaisi, Yael Lurie, Zaher S Azzam, Johad Khoury, et al. Analyzing medical research results based on synthetic data and their relation to real data results: systematic comparison from five observational studies. *JMIR medical informatics*, 8(2):e16492, 2020.

[14] Ashish Dandekar, Remmy AM Zen, and Stéphane Bressan. A comparative study of synthetic dataset generation techniques. In *International Conference on Database and Expert Systems Applications*, pages 387–395. Springer, 2018.

[15] Mohsen Dorodchi, Erfan Al-Hossami, Aileen Benedict, and Elise Demeter. Using synthetic data generators to promote open science in higher education learning analytics. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 4672–4675, 2019.

[16] Zahra Azizi, Chaoyi Zheng, Lucy Mosquera, Louise Pilote, and Khaled El Emam. Can synthetic data be a proxy for real clinical trial data? a validation study. *BMJ open*, 11(4):e043497, 2021.

[17] Robert Redpath and Bala Srinivasan. Criteria for a comparative study of visualization techniques in data mining. In *Intelligent Systems Design and Applications*, pages 609–620. Springer, 2003.

[18] Using Key Word Analysis of an Organization's Big Data For Error and Fraud Detection. (accessed on 8 September 2021).

[19] Emilie Lundin, Håkan Kvarnström, and Erland Jonsson. A synthetic fraud data generation methodology. In *International Conference on Information and Communications Security*, pages 265–277. Springer, 2002.

[20] Reverso Context. (accessed on 8 September 2021).

[21] Sentence Dict. (accessed on 8 September 2021).

[22] Random Word Generator. (accessed on 8 September 2021).

[23] Marco Sáncheza, Verónica Olmedo, Carlos Narvaeza, Myriam Hernándeza, and Luis Urquiza-Aguiar. Generation of a synthetic dataset for the study of fraud through deep learning techniques.

[24] Marco Sánchez-Aguayo, Luis Urquiza-Aguiar, and José Estrada-Jiménez. Predictive fraud analysis applying the fraud triangle theory through data mining techniques. *Applied Sciences*, 12(7):3382, 2022.

[25] Esra Kahya Ozyirmidokuz. Mining unstructured turkish economy news articles. *Procedia Economics and Finance*, 16:320–328, 2014.

[26] Nur Annisa Tresnasari, Teguh Bharata Adji, and Adhistya Erna Permanasari. Social-child-case document clustering based on topic modeling using latent dirichlet allocation. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 14(2):179–188, 2020.

[27] AUC. (accessed on 15 July 2021).

[28] Sirko Straube and Mario M Krell. How to evaluate an agent's behavior to infrequent events?—reliable performance estimation insensitive to class distribution. *Frontiers in computational neuroscience*, 8:43, 2014.

# 6 IMPROVING FRAUD DETECTION WITH SEMI-SUPERVISED TOPIC MODELING AND KEYWORD INTEGRATION

Marco Sánchez-Aguayo[1], Luis Urquiza-Aguiar[2]

[1]Departamento de Informática y Ciencias de la Computación, Escuela Politécnica Nacional, Ladrón de Guevara E11-253, Quito 170517, Ecuador

[2]Departamento de Electrónica, Telecomunicaciones y Redes de Información, Escuela Politécnica Nacional, Ladrón de Guevara E11-253, Quito 170517, Ecuador

Email: {marco.sanchez01, luis.urquiza}@epn.edu.ec

## 6.1 ABSTRACT

Fraud detection through auditors'manual review of accounting and financial records has traditionally relied on human experience and intuition. However, replicating this task using technological tools has represented a challenge for information security researchers. Natural language processing techniques, such as topic modeling, have been explored to extract information and categorize large sets of documents. Topic modeling, such as Dirichlet assignment (LDA) or nonnegative matrix factorization (NMF), has recently gained popularity for discovering thematic structures in text collections. However, unsupervised topic modeling may not always produce the best results for specific tasks, such as fraud detection. Therefore, in the present work, we propose to use semi-supervised topic modeling, which allows the incorporation of specific knowledge of the study domain through the use of keywords to learn latent topics related to fraud. By leveraging relevant keywords, our proposed approach aims to identify patterns related to the vertices of the fraud triangle theory, providing more consistent and interpretable results for fraud detection. The model's performance was evaluated by training with several datasets and testing it with another one that did not intervene in its training. The results showed efficient performance averages with a 7 % increase

in performance compared to a previous job. Overall, the study emphasizes the importance of deepening the analysis of fraud behaviors and proposing strategies to identify them proactively.

**KEY WORDS:** Fraud triangle; Human behavior; Topic modeling; Data mining; Text mining and Classification methods.



## 6.2 INTRODUCTION

Auditors can identify fraud by reviewing accounting and financial records; their experience can detect this phenomenon. Reproducing this task using technological tools has become a challenge for researchers in computer security. There are several initiatives to transfer auditors' knowledge to the technical area by applying machine learning techniques and theories related to fraud. However, representing intuitive expert judgments means a challenge, especially when the result of applying formal methodologies does not coincide with the experts'criteria. By applying topic modeling, it is possible to codify human knowledge and then use it to extract interpretable latent topics from a corpus. Topic modeling was introduced in an unsupervised environment [1], the most conventional being Dirichlet assignment (LDA) or non-negative matrix factorization (NMF), which have become popular in recent years. These are based on statistical models that allow discovering thematic structures in collections of texts, identifying themes humans can interpret and facilitating their understanding [2]. The structures found by unsupervised topic modeling often do not represent the best alternative for analyzing a specific phenomenon.

A topic modeler trained in reviewing documents in a corpus can be considered when this can discover the implicit semantic structures that describe general topics. However, we often want to dig deeper by discovering topics that reflect a specific behavior, in our case, related to fraud. Techniques for effectively finding patterns related to topics linked to a particular field are called semi-supervised topic modeling, which generates interpretable topics. For this reason, we propose using this modeling technique to learn latent issues about documents. Unlike LDA, these do not assume a specific data generation model and instead look for "more informative" topics.

In this context, we analyze in this article the most used semi-supervised topic models, such as Correlation Explanation (CorEx) [3] and Seeded LDA, to validate their performance in

a detector of suspected fraud behavior [4] that analyzes human behavior using the fraud triangle theory (FTT) [5] leveraged in machine learning (ML).

This will allow for flexibly incorporating the knowledge of the domain of study through the use of keywords within the topic modeling, which can lead the experimentation toward discovering topics that otherwise would remain hidden.

Using relevant words can help our proposed detector recognize patterns related to the vertices of the fraud triangle, thus allowing the analysis to be guided directly on this fraud theory.



### 6.2.1 Related Work

Semi-supervised topic modeling has been the focus of various research studies in domains such as clinical notes and marketing. These studies have proven to be valuable as they offer topics that are easily interpretable. For example, [6] utilized the Anchored Correlation Explanation (CorEx) algorithm to extract English tweets related to eating disorders, aiming to develop a tool for identifying this disorder. Another study by [7] introduced an automated medical image retrieval system incorporating subject and location probabilities to enhance performance. Using the guided latent Dirichlet assignment (GuidedLDA) method facilitated the generation of topic information. This approach demonstrated superior average mean precision (86.74) and precision (97.5) compared to previous methods.

[8] suggested using topic modeling to identify human factors-related topics in aviation safety reports. Utilizing algorithms like CorEx and SeededLDA achieved more accurate results without requiring manual revisions. Similarly, [9] also explored using the CorEx algorithm for topic modeling, aiming to extract interpretable latent topics by harnessing informal human knowledge. The study by [10] employed various topic modeling techniques to analyze chat data collected in a library to extract specific and easily interpretable topics. They evaluated the results quantitatively using the coherence metric, while a librarian who was also an author of the article assessed qualitative accuracy and interpretability. [11] presented a topic-modeling approach incorporating relevant words to identify rare diseases not mentioned in clinical health notes. The objective was to provide relief workers with better guidance in offering practical help and eliminating ambiguities when analyzing complex problems.

In a recent study, [12] utilized topic modeling to evaluate semantic relationships in short messages on Twitter. They could identify associations with specific discussion topics by analy-

zing the hashtags used in these messages. This method proved helpful in understanding the content and context of these messages. [13] took a qualitative approach and developed a natural language processing tool called Guided Latent Dirichlet Allocation (GLDA). This tool analyzed entertainment products, such as award-winning films, based on media psychology literature. By predicting viewers' behavior, they demonstrated the potential of this approach for understanding consumer behavior in film selection. [14] focused on generating new document classification systems using automatic learning methods. They employed LDA to identify groups of words related to the attributes of the documents, enabling efficient document search based on matching keywords by topic.

To address the issue of overlapping topics, they utilized guided LDA, which allowed them to influence topic generation by setting seed words per topic. Another study by [15] proposed a method to identify seed words for disaster-related topics automatically. By comparing words from tweets on the day of the disaster occurrence with the previous day in the same area, they could obtain initial words using LDA. These words were then used to identify tweets related to the event. This method proved effective in automatically identifying relevant words for disaster-related topics. [16] evaluated different topic modeling algorithms for knowledge extraction in the tourism industry. Their findings showed the complexity of analyzing short-text social media data and emphasized the effectiveness of using CorEx to analyze Instagram content. CorEx outperformed LDA and NMF in ranking relevant sites and activities. LDA results were homogeneous and overlapping, while topics extracted from NMF were not specific enough to gain deep insights.

These research works demonstrate the diverse applications and benefits of semi-supervised topic modeling in different domains. Using algorithms like CorEx and GuidedLDA allows for more precise and interpretable topic extraction. This enhances our understanding of complex topics and enables the development of practical tools for identifying specific disorders, improving medical image retrieval systems, and analyzing human factors in safety reports.

Additionally, Table 6.1 presents a summary providing information including methods used, publication year, fields, and purpose to the significant state of the art.

| Topic | Field | Authors | Method Used | Purpose/Outcome |
|---|---|---|---|---|
| Medical and Healthcare | Clinical Notes | - | Semi-supervised Topic Modeling | Extract interpretable topics |
| | Eating Disorders | Pecore et al. (2021) | Anchored CorEx | Identify eating disorders |
| | Medical Image Retrieval | SHAMNA et al. (2019) | GuidedLDA | Improve image retrieval |
| | Rare Disease Recognition | Gallagher et al. (2016) | Topic Modeling | Recognize rare diseases |
| Human Factors and Safety | Aviation Safety | Lyall-Wilson (2019) | CorEx, SeededLDA | Identify human factor-related topics |
| Social Media Analysis | Human Knowledge | Reing et al. (2016) | CorEx | Extract informal human knowledge |
| | Semantic Relationships | Steuber et al. (2020) | Topic Modeling | Analyze semantic relationships |
| Cultural and Entertainment | Entertainment Description | Toubia et al. (2018) | Guided LDA | Analyze films and predict behavior |
| Information Retrieval | Document Classification | Hoffmann (2021) | Topic Models with Metadata | Enable document search using topics |
| Disaster and Event Related | Disaster Identification | [15] | LDA | Identify disaster-related topics |
| Tourism and Social Media | Tourism Knowledge | Egger et al. (2021) | Topic Modeling | Analyze tourism content |
| Library and Information | Library Chats | Koh & Fienup (2021) | Various Topic Modeling | Analyze library chat data |

**Table 6.1:** Research Papers Grouped by Topics and Fields

A detailed study on fraud-related jobs was conducted in [17]. A Systematic Literature Review (SLR) proposes collecting and analyzing research that addresses this phenomenon, considering human behavior as the leading risk factor reviewed associated theories that study this phenomenon. In addition, Machine Learning techniques were incorporated into the research that allows their detection.

This work was developed in the context of a previous investigation entitled “Predictive Fraud Analysis Applying the FTT through Data Mining Technique” [4]. They propose a detector of suspected fraud behavior by analyzing human behavior using the FTT leveraged in machine learning (ML) and deep learning (DL). To develop this proposal, they evaluated the performance of frequently used text mining techniques, such as LDA, NMF, and latent semantic analysis (LSA). Finally, to determine the differences in performance, they used receiver operating characteristic (ROC) curves based on the area under the curve (AUC) with the traditional ML classification methods to identify which technique is more compatible with the modeling of topics to detect suspicious behavior of fraud.

In this context, the present work proposes to deepen the analysis of topic modeling through the use of semi-supervised techniques associated with fraud theories that, through classification algorithms, make it possible to more efficiently detect possible cases of fraud not observed in the works mentioned above. Therefore, this represents a clear research gap in this area.



### 6.2.2 Contribution

The main contributions are the following: first, we use CorEx as a topic model and perform an efficient alteration of its code to identify the probabilities that the corpus documents belong to a topic and to be able to visualize the distribution of topics through the pyLDAvis library. Second, we show how the FTT can be integrated into CorEx through “keyword” related to the vertices of this theory. We show that CorEx produces more relevant topics than its unsupervised and semi-supervised variants of LDA.

Once the most efficient semi-supervised topic modeling has been identified, the probabilities that a document belongs to a specific topic are obtained, with which classification methods such as Gradient Boosting (GB) and Random Forest (RF) were trained to try to predict related cases with fraud. Finally, the proposed model is validated with the different datasets used in this research to try to establish the generality of the model.

Several synthetic datasets were used and generated to validate their performance to ensure the model’s accuracy. The datasets were generated using various techniques to simulate different scenarios and environments. The model was tested in multiple conditions to ensure it worked reliably in all situations, confirming that it could accurately predict outcomes in va-

rious contexts. The results of these tests were then used to validate the model's performance and provide evidence of its accuracy.

The rest of this document is organized as follows. The "Background" section provides relevant information on FTT, topic modeling, and machine learning classification methods. Then, the Section "Methodology" describes the data preparation and the methodology used in this work. Next, the Section "Results and Discussion" deals with the experiment, the results, the validation, and the discussion. Finally, the "Conclusions" section is presented, addressing future work.



## 6.3 MATERIALS AND METHODS

This section briefly describes the fraud triangle theory, topic modeling strategy, classification methods, and validation methods.

### 6.3.1 Fraud Theories

Today's society is constantly changing due to factors like globalization, technological advancements, and the rapid growth of industries. This creates several difficulties, particularly those about information security and management. Due to this, there may be an increase in fraud risk for both public and private companies. Organizations are now more aware of the need for fraud detection and prevention techniques due to the high crime rates to reduce the risk of fraud. [18]. Organizations face a severe problem with cybersecurity and the risks that come with it due to internal and external factors worldwide. The internal ones are related to the companiesínherent management and commercial activity, while the external ones are related to politics and the global economy. These risks exist, increasing the chance that they could become a fraud [19]. The Association of Certified Fraud Examiners (ACFE) classifies occupational fraud into three types: asset misappropriation, corruption, and fraudulent statements. Asset misappropriation refers to the theft or misuse of an organization's assets. Corruption influences a business transaction for personal gain, and misrepresentation is the intentional misrepresentation of financial or non-financial information to deceive others [20]. Several theories allow analyzing the problems related to fraud, which serve as a guide for organizations to combat this phenomenon, contributing to the prevention, detection, and deterrence of activities related to occupational fraud. Why is labor fraud committed within

organizations? This question explains the fraud triangle, the first model developed to address this problem. This theory has been the basis for creating tools to deal with this crime. However, it has its limitations, which do not cover all fraud cases due to the progress and sophistication of this behavior, so developing a model that includes all fraud cases is a challenge [21] [22]. Over the past 60 years, the fraud triangle has evolved into various models, including the diamond and the fraud pentagon. The FTT was proposed by Cressey (1953); FTT identifies three crucial elements: pressure, opportunity, and rationalization. According to this theory, fraud is typically accompanied by pressures/incentives, opportunities, and rationalizations/attitudes. Thus, it is highly probable that the perpetrator is driven by pressure or motivation to commit fraud. Additionally, the perpetrator will likely find potential opportunities to carry out their fraudulent actions. Moreover, they can rationalize their deceitful acts by justifying their necessity. Ultimately, all three conditions directly correlate with a heightened likelihood of fraud [23]. This theory later evolved into the fraud diamond theory by adding a new element, capacity, proposed by Wolfe and Hermanson (2004). Finally, the Pentagon theory of fraud is the latest evolution proposed by Jonathan Marks (2012), to which two elements were added: competition and arrogance. Competition in this model has the same meaning as the ability described by Wolfe and Hermanson in 2004, aiming to perfect the diamond theory of fraud [24]. The elements or variables associated with the different fraud theories are directly related to the behavior of the perpetrators, which are clear indicators that can cause fraud. The triangle, diamond, and pentagon of fraud are relevant theories that can be used interchangeably effectively to detect the possibility of fraud, depending on the existence and availability of evidence related to the variables of these theories [25]. The effectiveness of the fraud triangle theory has been proven in [26], evidencing more precise results on the fraud diamond and pentagon because the capacity and arrogance of the variable in many cases do not significantly affect the behavior of fraudsters. Individuals will not commit fraud despite great ability and arrogance. In this context and because the characteristics of the study dataset are aligned with the triangle theory of fraud, this model will be used to develop this work.



### 6.3.2 Topic Modeling (TM)

Topic modeling is a statistical technique that has revolutionized text mining, allowing the discovery of semantic structures in a collection of documents [27]. Popular algorithms for multi-

domain text analysis include Latent Semantic Analysis (LSA), Non-Negative Matrix Factorization (NMF), Probabilistic Latent Semantic Analysis (PLSA), and Latent Dirichlet Allocation (LDA). LSA and NMF work on a bag-of-words (BoW) model-oriented approach, a text representation describing the occurrence of words within a document, which converts a corpus into an array of document terms. On the other hand, LDA and PLSA were initially unsupervised approaches, which evolved into supervised and semi-supervised models, respectively [28]. These models have weaknesses associated with their design; in the case of LSA, obtaining and determining the optimal number of topics is a complex task. PLSA has several overfitting problems, and LDA often does not expose the relationships between topics. To circumvent these difficulties, topic modeling with a semi-supervised approach allows previous knowledge to be provided in the topic model. Specifically, there are versions in which the model can be given "seed" words of the study topic, and the model's algorithm encourages topics to be built around these seed words; this solves the problems mentioned above and allows us to direct topics toward relevant topics simply by adding keywords while leaving room for discovering "unknown" topics. In this context, alternative models to the traditional techniques have been developed in the semi-supervised approach, such as Correlation Explanation (CorEx), which, unlike LDA, does not make assumptions about the data generation process but instead addresses the modeling of issues.



In an information-theoretic way, they avoid time and effort to identify topics and their structure ahead of time. On the other hand, guided LDA (GuidedLDA), a variant of LDA, improves the performance of topics that infrequently occur, where a variation of the LDA algorithm is made so that the topic-word and topic-document distributions take into account the seed words [29]. They have also appreciated techniques such as the Dirichlet multinomial mixture (DMM) that allows for overcoming data scarcity problems in short texts, generally below 500 characters [30].

#### 6.3.2.1 GuidedLDA

GuidedLDA or SeededLDA implements LDA and can be guided by setting some seed words per topic, which will cause topics to converge in that direction [31]. In the study by [32], they used words that belong to specific topics and are limited to appearing in some subset of all possible topics. A second model proposed by [33] uses relationships between words to break up confusing topics. While in [34], they propose SeededLDA, an extension

of semi-supervised LDA, and use seed words to influence both topic-word distributions and document-topic distributions; it is a model that guides but does not force these topics into seed words. Specifically, the generative process of estimating these distributions is guided by initial word-level information, a set of user-defined words characteristic of the topics in the study corpus. This approach allows the user to provide N sets of representative seed words from the corpus to guide the topic discovery process. These "seed sets" correspond to the word sets preliminarily obtained by the LDA [35] model. To obtain contextually relevant topics, such as the impacts of fraud, strategies, and initiatives for its prevention and mitigation, some initial keywords highlighted by topic must be established, allowing us to obtain topics that help us understand the content of the dataset we are analyzing.



#### 6.3.2.2 (Anchored) Correlation Explanation (CorEx)

CorEx is a topic model based on Total Correlation Explanation, which allows identifying topics in a corpus and explaining their structure through the dependency found on the data [36]. In addition, it is a seeded technique with several advantages over the seeded LDA variant (SeededLDA), such as a better consistency of the derived topics and good algorithmic performance [37]. Unlike LDA, CorEx makes no assumptions about the data generation process but instead approaches topic modeling in an information-theoretic way. This model allows the incorporation of domain knowledge through user-specific anchor words that guide the model to topics of interest; this allows the model to represent topics that do not arise naturally and provides the ability to separate keywords that allow topics to be identified differently [38]. Anchored CorEx optimizes the following in Equation (6.1) [39]:

$$\underset{X;Y}{\text{Maximize}} \quad TC(X;Y) + \beta \sum I(x;y) \tag{6.1}$$

Where X and Y are random variables, TC and I represent the total correlation and mutual information, respectively, and x is an anchor word.

### 6.3.3 Classification methods

Classification problems have been deeply analyzed and have aroused the scientific community's interest, mainly applied in data analysis in machine learning, statistical inference, and

data mining [40][41]. In general, classification is a data mining approach used to predict the membership of a data instance to a given class from a set of predefined classes [42][43].

Given such a diversity of methods, the question arises as to which method should be used for a problem to be solved. The answer depends on the nature and approach with which the problem is addressed. So there will be many performance measures, each addressing different aspects [44].

Using the AUC criterion, this paper compares RF and GB to detect fraud-related text. A description of these algorithms can be seen in Table 6.2.



**Table 6.2:** Description of two classification methods: Random Forest (RF) and Gradient Boosting Decision Tree (GBDT).

| Model | Description | References |
|---|---|---|
| Random Forest (RF) | A tree-based ensemble where a set of random variables determines each tree. Decision trees are chosen randomly from the available data, and the averaging process helps mitigate low bias and high variance. | [45, 46, 47] |
| Gradient Boosting Decision Tree (GBDT) | Use decision trees as weak classifiers for regression or classification tasks with logarithmic loss. It combines the results of multiple variables sequentially to outperform earlier outcomes by using gradient increase to train predictors and repair previous mistakes. | [48, 49, 50] |

Frequently, the performance of a combination of indicators is quantified by indices related to the Receiver Operating Characteristic (ROC) curve: sensitivity, specificity, or the area under the curve (AUC) [51]. A ROC curve is a graph that shows the relationship between the true positive rate (TPR, or specificity) on the y-axis and the false positive rate (FPR, or 1 – specificity) on the x-axis [52]. The ROC curve shows the performance of a classifier without considering the class distribution or the cost of misclassification. The area under the receiver operating characteristic curve (AUC) must be determined to compare the ROC curves of various classifiers [53]. The area under the ROC curve, or AUC, measures model performance for all possible decision thresholds. It gauges the overall performance of a test set and is interpreted as the average sensitivity value for all potential specificity values. Since the x and y axes have 0 to 1, it can take any value between 0 and 1 [54].

## 6.4 METHODOLOGY FOR PREDICTING FRAUD BASED ON THE FRAUD TRIANGLE COMPONENTS



Implementing a predictive model that identifies hidden patterns related to suspected fraud is the objective of this work, for which topic modeling was used and, specifically, the most relevant techniques used in text mining, such as LSA, NMF, and LDA, were tested. A comparison was made to identify these algorithmséfficiency and determined that LDA is the most consistent model. To determine the number of topics, the coherence value or parameter k was used as a metric, which allows us to identify the most appropriate number of topics of the three models that adjusts to the nature of the information used and indicates the level of similarity. Semantics exist between words for each topic. LDA allows finding topics to which a document belongs based on the words it contains. This served as a starting point to identify the most representative words and their distribution in the different topics. This initial strategy served as a starting point for using semi-supervised learning algorithms by using some initial words for the topics considered most representative of the underlying themes in the study corpus. It guided the models to converge around those terms. This way, we observe how the models can configure the seed words to guide their results in a particular direction.

The application of topic modeling aims to determine the probability that a document within the study corpus belongs to a specific topic that aligns with the vertices of the fraud triangle. This crucial step, depicted in the first phase of Figure 6.1, identifies potential fraud-related behaviors. These probabilities are then used to train various classification methods, allowing for predicting suspicious activity associated with fraud. Evaluating the performance of the different classifiers is essential in selecting the one most compatible with the topic analysis carried out for fraud detection. This fundamental evaluation stage, illustrated in the second phase of Figure 6.1, ensures the effectiveness and accuracy of the chosen classifier.

### 6.4.1 Dataset generation

One of the most difficult challenges in the analysis and study of fraud is the lack of information related to this phenomenon. The datasets that contain evidence that identifies fraudulent activities or suspicions of possible commissions of this crime are scarce and difficult to access due to their confidentiality, rights, and intellectual property. Due to these difficulties, a

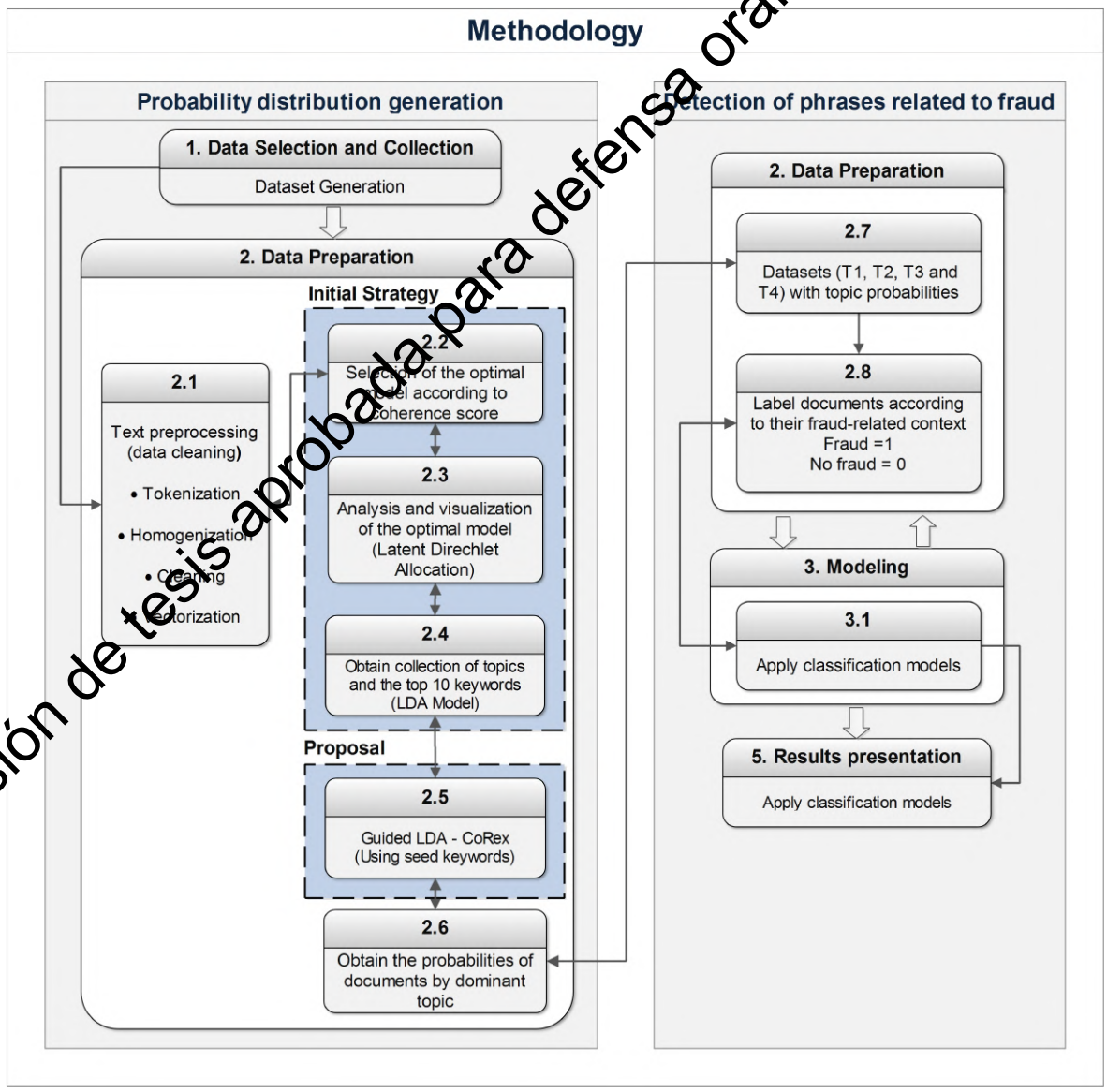




**Figure 6.1:** Methodology used to determine the existence of fraud.

practical solution that solves this need is to generate synthetic data, which becomes a viable strategy for studying this phenomenon. According to several studies, this data type allows experimentation using machine learning techniques to be faster and more efficient, producing data with similar characteristics to reserved and difficult-to-access information [55].

In [4], they used a synthetic dataset generated from a dictionary of keywords related to the fraud triangle, which we will call WebScraping. Through the use of online tools [56] that allow the production of grammatically well-defined sentences from the entry of specific words, which for this particular case were used those belonging to the dictionary of words related to fraud and its vertices (pressure, opportunity, and rationalization), which served to build the study dataset in this research. This study proposed an initial fraud prediction model

supported by the synthetically generated dataset. To validate the usability of synthetically generated data, [57] compares and evaluates the performance of different synthetic datasets (WebScraping and Neural-Network) versus an original (Students) to demonstrate whether the synthetic data can be used as a substitute for real data, noting that according to the performance metrics obtained from the comparison made, this alternative to real data is reliable and serves as a valid option for data analysis. The experimentation on the real and synthetic datasets allows the identification of similar behaviors in the results based on their performance after applying the fraud prediction method, which suggests that the different datasets analyzed can be generalized to different scenarios. In this context, the datasets mentioned above were used for the present work, with the objective of contrasting results after applying modeling of semi-supervised topics and classification algorithms vs. the first fraud prediction model proposed.

A dataset called ChatGPT was also generated, with the same characteristics as the previous ones, which this tool was used to build. This artificial intelligence made it possible to generate phrases related to the three vertices of the fraud triangle, for which messages were entered that consisted of imagining scenarios that include the elements of the theory; in the case of opportunity, they were told to “imagine a scenario in which that an employee has access to confidential financial information,” to incorporate the element of rationalization, he was asked to “imagine a scenario in which an employee justifies his fraudulent actions by believing that he is not being fairly compensated” and finally, to incorporate the pressure element, “Imagine a scenario where an employee experiences financial hardship that motivates them to commit fraud.” Once this environment was created, the generation of a certain number of sentences oriented to each vertex of the triangle was promptly requested, entering messages requesting “Tell me sentences that a person can say regarding pressure, opportunity, or perceived rationalization”



### 6.4.2 Data preprocessing

Within Artificial Intelligence, natural language processing (NLP) is the field that is responsible for investigating how computers understand, analyze, and interpret human language. NLP allows people to interact with machines in human language. Computer languages only work correctly when correctly written because they are precise in their syntax. At the same time, the flexibility of natural languages allows them to adapt to their nature and interpret errors

such as accents, words, and dialects [58]. One of the most common NLP tasks is to clean up text data. Extracting the text to the most critical root words in the corpus maximizes the results. Text preprocessing in NLP is a method that allows cleaning up text so that it is ready to feed models. Noise in the text comes in various forms, like punctuation and different cases. All these noises are not helpful for the machines and, therefore, need to be cleaned [59]. In text mining, preprocessing involves a 3-step mechanism that includes extraction, stopword removal, and lemmatization. Extraction is the process of breaking down documents into individual elements, formed a format composed of tokens, words, terms, or attributes. These features represent the document in a vector space, with their weights determined by the frequency in the text document. Removing stopwords, numbers, and special characters helps reduce the dimensionality of the term space. Lastly, lemmatization standardizes words by reducing them to their etymological root, eliminating common suffixes, and reducing word count [60].



### 6.4.3 Quantitative evaluation of topic modeling algorithms

The structured grid search technique was used to identify the optimal topic modeling. The results obtained by the unsupervised LDA algorithm were compared with semi-supervised models such as CorEx and guided LDA. In related works, comparisons between these techniques are presented, evidencing, in most cases, the superior performance of semi-supervised algorithms over unsupervised ones [61] [62]. In this context, there is little evidence of studies comparing semi-supervised models. It is necessary to analyze the efficiency of these semi-supervised algorithms, for which, in the first instance, the same text preprocessing techniques and the hyperparameters used as input for the models were used, and they were evaluated through the coherence value “C_v” That determines the performance of the algorithms the different topic modeling algorithms. This metric allows us to identify how coherent a model is about the structure of its topics; the more different the words are in each topic, the less related the topics will be, and the more coherent the model will be. Once the hyperparameter k or an adequate number of topics was identified to obtain adequate modeling, the models were tested using seed word dictionaries to more easily generate topics corresponding to the identified categories.

### 6.4.4 Selection of the topic modeling algorithm

Once the models generated by GuidedLDA and CorEx were obtained, the consistency of the sets of words per topic formed was analyzed. The efficiency of the semi-supervised algorithms to establish the distributions in each topic was determined. This analysis allows us to identify, according to the parameters, the most suitable method to distribute the words in their respective topics more efficiently. With the different modeling results, we analyze the word distributions by topic and identify the words related to fraud and their behavior within the distribution. We point out those that coincide with the seed word dictionaries used, which are associated with the vertices of the fraud triangle. The objective is to show if the topics generated are associated or related to the vertices of the pressure, opportunity, and rationalization triangle. After this analysis, we select the model with the best performance and that most consistently brings together the words related to fraud by topic, through which we will obtain the probabilities that the documents in the study dataset belong to a given topic. The different probabilities obtained represent a measure that makes it possible to identify whether a document is related to fraud and express a new representation of the dataset. Then, we build smaller datasets from this new dataset, each of which groups documents associated with a “dominant” topic, a topic to which the documents most likely belong.



### 6.4.5 Evaluation

Once the appropriate topic modeling for the present case study has been identified and the probability distributions of documents per topic have been generated, it is feasible to use machine learning techniques to predict fraud-related activities. When small datasets are available, traditional classifiers frequently learn better than deep learning classifiers, which gives us a guideline for selecting the appropriate techniques. The graph of the ROC curve and its area under the AUC curve was used to evaluate the performance and identify how accurate the prediction of the classification methods used in the experiment, which represents the quality of the methods, which allows us to visualize the behavior of each of these and analyze their performance.

In addition, as part of the evaluation process, training on the proposed model will be carried out using datasets generated one at a time. Later, it will be tested with the remaining datasets; this will allow for obtaining more accurate and reliable results on the effectiveness and

performance of the model in different scenarios, guaranteeing a comprehensive evaluation of its performance under different conditions.

## 6.5 RESULTS AND DISCUSSION

This section presents the results obtained from testing our improved fraud detection model. The efficiency of the results is analyzed and discussed in this section. Details about the experimentation, from selecting topic modeling to applying machine learning models, are reviewed. Finally, the different theories, techniques, and models applied to the approach of this model are discussed.

### 6.5.1 Probability distribution generation

The first stage of the experimentation consists of applying topic modeling techniques to the study dataset to identify hidden patterns related or not to fraud and analyze how consistent these results are. This is to obtain information structured by topics that, once the model is applied, allows us to analyze its characteristics based on the probability that a document belongs to a specific topic.

#### 6.5.1.1 Initial Strategy - Application of the LDA Model

Of the topic modeling algorithms reviewed in [4], it was determined that LDA has the best behavior when analyzing data related to fraud since it more consistently groups words by topic. After carrying out different tests in the experimentation, it was validated that the adequate number of topics is 4. With this value, the LDA algorithm is applied to the study dataset, obtaining a distribution of words categorized into four topics according to their context and problem of study. The present work relates to the vertices of the FTT, “pressure, opportunity, and rationalization,” and another topic they call others. This distribution of words can be seen in Table 6.3, which is ordered by topic and prevalence. In addition, the words related to fraud are colored to identify that they belong to a specific vertex of the fraud triangle. Words unrelated to fraud that belong to said vertices were not colored. The top 20 terms are manually analyzed and filtered to use only the most significant ones (for each topic).

Table 6.3 presents inductive labels, presenting the main terms identified by the class and the most significant to be used as seed or anchor words for the semi-supervised models, which are identified by colors. For example, in Topic 2 (T2), words like “Life” or “word” are related to a different approach to fraud, and, therefore, we did not choose them as meaningful representations for one of the vertices of the fraud triangle. As can be seen, the words related to fraud are distributed through the topics without distinguishing groups in the different topics; this indicates that the topics obtained through LDA cannot be directly associated with the vertices of the fraud triangle. However, due to the presence of a high number of words with a high degree of repetition in the different topics, the existence of behavior related to fraud follows.



**Table 6.3:** The most frequent terms in the dataset connected to each of the three vertices of the fraud triangle are found after LDA has been applied. To represent the vertices of pressure, rationalization, and opportunity, the words are colored orange, blue, and green.

| Topics | | | |
|---|---|---|---|
| **T1** | **T2** | **T3** | **T4** |
| review | debt | problem | want |
| care | think | economic | know |
| poor | later | make | job |
| steal | fix | big | work |
| temporary | just | people | lose |
| say | tell | abuse | support |
| new | inadequate | fair | deadline |
| man | look | compensation | help |
| really | failure | child | come |
| insufficient | weakness | good | time |
| state | ill | earning | exploitation |
| money | unethical | easily | deserve |
| issue | life | accessible | scare |
| evacuation | world | country | right |
| leave | try | need | like |
| woman | let | way | day |
| year | talk | pay | use |
| long | old | school | scared |
| change | feel | home | ask |
| period | place | thing | car |

#### 6.5.1.2 Proposal - Explore topic modeling using semi-supervised learning

Classical topic modeling methods are algorithms that generate various topics from a study dataset. However, due to their unsupervised nature, these methods can impede the comprehension of the analyzed texts. They are prone to create less essential topics, leaving aside several others that may interest them [63]. Each word is randomly assigned to a topic in LDA, controlled by Dirichlet priorities through the Alpha parameter. Then, it is required to determine which term belongs to which topic. LDA uses a straightforward approach to fin-

ding the topic for one term at a time. Suppose we want to find the topic of the word “problem” related to fraud. The algorithm distributes each word evenly across all the topics found and assumes it is the right topic for those words. Then, find out what other words the word “problem” is associated with most often. In this context, it is determined what the most common topic among those terms is. Therefore, the word “problem” is assigned to that topic. The word “problem” is close to any topic where words like “debt” and “need” are found. These three words are closer to each other before this step. Finally, the model moves on to the next word and repeats the process as many times as necessary to converge. Semi-supervised topic modeling allows the introduction of prior knowledge by incorporating words called “seed” or “anchor” into the algorithm that stimulates or encourages (but does not force) the model to build topics around these anchor words. This alternative of adding keywords gives the flexibility to generate relevant topics while allowing the discovery of unknown topics. GuidedLDA and CorEx use tag seed words to make their training converge around these words. That is, a set of specific words relevant to a tag related to the same topic is used, and the weight of these particular words is increased during training to capture other strongly related words. In other words, these seed words function as anchors.



The coherence score aims to measure the similarity between words and how interpretable the topics obtained by the model are. Starting from the premise that we have the reference coherence score obtained for the LDA model, in which several sensitivity tests were carried out to determine the adequate number of themes, they established C_v as a metric for performance comparison. Since the coherence score gradually increased with the number of topics, the model with the highest C_v was chosen. In this case, K=4. For the semi-supervised models that we will use in this proposal and considering that we will use the same study corpus, we will use this value of K-topics to perform the respective tests.

As mentioned, semi-supervised topic models require a list or set of keywords called seed or anchor related to each topic for modeling. These words are used to identify specific topics; in this sense, they are related to the three vertices of the fraud triangle theory for the present work. In these models, a force or push parameter defines the bias of the generated topics toward the seed or anchor keywords. This value can vary between models; for the case of GuidedLDA, it can range between 0 and 1. A 0.1 can bias the seed words by 10 % more toward the seed topics. On the other hand, in CorEx, it should always be above 1, and higher values indicate a more substantial bias toward anchor keywords. In this context, the list of anchor keywords for the models was provided, those that were generated in the initial

strategy applying LDA, represented in Table 6.3, and those words with the greatest representativeness related to the three vertices were chosen from the different topics “pressure, opportunity, and rationalization.”

Words are initialized by setting tags as keys and a list of initial words (relative to critical topics) as values.

```
keywords=[
        [économic', 'problem', 'deadline', 'review', 'debt', éxploitation',
        'lose', 'job', 'scared'],
        [éarning', 'insufficient', ínadequate',évacuation', 'supervision','weakness',
        érror', 'failure', 'support', éasily'],
        ['deserve', ábuse', 'fair', 'temporary', únethical', 'poor',
        'steal', 'care', 'fix', 'later'],
        ['love', 'study', 'think', 'time', 'people', 'write',
        'play', 'game', 'passion']
        ]
```



Table 6.4 shows the results of applying GuidedLDA and CorEx topic modeling and the top 20 terms identified by topic. It can be observed how the words of the study corpus are distributed in the four established topics. These words are organized by topic and prevalence to identify those that the model considers most relevant. Using the same procedure as [4], we color the words to identify their belonging to each vertex of the fraud triangle. Those not related to the vertices were not colored. We can observe that the model obtained by GuidedLDA has a behavior similar to that of regular LDA, which does not reflect a relationship between the topics obtained with each of the vertices of the fraud triangle since the words within each topic contain words associated with different vertices of the triangle. As in regular LDA, the model does not group words into topics related to each vertex of the fraud triangle. Still, the probability that the corpus documents belong to each topic provided by the model is helpful for feed classification algorithms to detect whether or not a phrase is related to fraud. On the other hand, the results obtained by CorEx are more interpretable than their predecessor since each topic obtained can be linked directly with the knowledge of the domain established in the list of initial words or anchors. A clear relationship can be seen between the resulting topics and the vertices of the fraud triangle; for example, in topic 1 (T1-CorEx), the words (orange) related to pressure are grouped in order of importance; in topic 2 (T2-CorEx), the words (green) related with opportunity, topic 3 (T3-CorEx) the words

(purple) related to rationalization and finally topic four those that are not related to fraud. CorEx allows the obtained model to converge to link the seed or anchor words to a given topic.

**Table 6.4:** The terms that appear most frequently in the study dataset are associated with each of the three vertices of the fraud triangle once GuidedLDA and CorEx have been applied. The words are colored orange, blue, and green to indicate the vertices of pressure, rationalization, and opportunity, respectively. CorEx better classifies the terms by topic.

| Topics | | | | | | | |
|---|---|---|---|---|---|---|---|
| **T1** | | **T2** | | **T3** | | **T4** | |
| G-Lda | CorEx | G-Lda | CorEx | G-Lda | CorEx | G-Lda | CorEx |
| review | problem | love | support | people | care | think | people |
| debt | economic | system | failure | study | poor | write | think |
| economic | review | love | easily | think | deserve | lose | time |
| study | job | failure | insufficient | play | later | people | love |
| problem | lose | study | inadequate | abuse | compensation | nobody | play |
| deadline | deadline | write | evacuation | economic | fix | care | privacy |
| earnings | exploitation | play | earning | accessible | steal | deserve | tank |
| compensation | labor | error | supervision | poor | temporary | steal | song |
| inadequate | period | fix | error | exploitation | fair | job | album |
| fair | currently | weakness | accessible | problem | unethical | play | update |
| insufficient | solve | evacuation | security | many | illegal | poor | indigenous |
| problems | social | accessible | muscle | unethical | trade | fix | spend |
| play | political | think | file | supervision | seek | something | change |
| supervision | issue | temporary | datum | temporary | know | unethical | live |
| exploitation | country | job | strength | role | alcohol | want | make |
| countries | work | file | remain | problems | victim | song | people |
| role | external | use | capacity | children | try | things | think |
| period | face | case | warn | love | verbal | good | time |
| people | | data | | weakness | | anything | |
| evacuation | | change | | work | | look | |

Once the models are obtained, the four resulting topics are manually labeled concerning the three vertices of the fraud triangle: pressure, opportunity, rationalization, and others. This categorization of topics is essential since it allows for interpreting the corpus and identifying the implicit topics in a dataset. The interpretation of a topic can be achieved by examining a ranked list of terms in each topic [64].

With the defined models, we can obtain the probability that a particular document in our corpus belongs to a specific topic; by entering the document into the model, it analyzes it and calculates the possibility of belonging to each one of the topics, establishing a percentage of probability per topic. One approach to classifying a document as belonging to a particular topic is to analyze which topic contributed the most to that document and assign it to that topic. Table 6.5 shows the percentages of belonging to a document associated with each topic. In this case, the one with the highest value corresponds to the most related or dominant for GuidedLDA.

Applying the same procedure to CorEx, it can be seen in Table 6.5 that the algorithm returns boolean values (True or False) to determine if a topic contributed more to a document and

| Docs | Pressure | | Opportunity | | Rationalization | | Others | |
|---|---|---|---|---|---|---|---|---|
| | G-LDA | CorEx | G-LDA | CorEx | G-LDA | CorEx | G-LDA | CorEx |
| 0 | 0.43 | False | 0.12 | True | 0.45 | False | 0.00 | False |
| 1 | 1.00 | False | 0.00 | False | 0.00 | False | 0.00 | False |
| 2 | 1.00 | False | 0.00 | False | 0.00 | False | 0.00 | False |
| 3 | 0.01 | False | 0.00 | True | 0.98 | False | 0.01 | False |
| 4 | 0.00 | False | 0.00 | False | 0.23 | False | 0.77 | True |
| 5 | 0.99 | False | 0.00 | True | 0.00 | True | 0.01 | False |
| 6 | 1.00 | True | 0.00 | False | 0.00 | False | 0.00 | False |
| 7 | 0.23 | True | 0.00 | False | 0.00 | False | 0.77 | False |
| 8 | 0.99 | False | 0.00 | True | 0.01 | False | 0.00 | False |
| 9 | 1.00 | False | 0.00 | True | 0.00 | False | 0.00 | False |

**Table 6.5:** Probabilities obtained from GuidedLDA (G-LDA) and CorEx in the different established topics; where each row represents a specific result for a particular model, the values in the G-LDA column represent the probability obtained by this model that a document belongs to that topic. In contrast, the values in the CorEx column have binary values, where true indicates that the document belongs to that category, and false indicates what is contrary.



if this document is more related to that topic. In general, this approach could inform about a document belonging to a specific topic without specifying the weights to which each topic contributed to that document.

Given that the metrics corresponding to the probabilities obtained by the models are necessary to feed classification models and their subsequent fraud prediction, it is essential to obtain the required values and be able to process them using machine learning algorithms.

In this context, the operation of the “corextopic.py” module, developed in Python, which contains the functions associated with transforming the data according to the previously defined model, was analyzed. The transform() algorithm 4 takes a matrix X consisting of (n_samples, n_visible), where n_samples is the number of data points, and n_visible is the dimensionality of each data point. The input data samples are preprocessed by applying a normalization or standardization, which is done by calling the preprocessing method; the preprocessed data is then stored in X. The latent_calculation method is then called to calculate the latent variables p(y|x) and the log-likelihood of the data log(z) for the preprocessed data samples X and the model parameters (self.theta). The resulting values are stored in p_y_give_x and log_z, respectively. Finally, the label() method is called to assign a label to each data sample; this algorithm 4 is inside the same “corextopic.py” class, which takes the matrix p_y_given_x of the form [n_samples, n_hidden] that represents the distribution over the hidden variables given the observed variables and returns binary labels for each sample based on the estimate of maximum likelihood. Additionally, it applies a threshold of 0.5 to the probabilities at p_y_given_x. If the probability of the hidden variable is more significant than 0.5, the method

assigns it a true label; otherwise, it assigns a false label. The output is a boolean array of the form [n_samples, n_hidden] representing the labels of each sample. The resulting labels are stored in the labels variable.

**Algorithm 4** Label hidden factors for (possibly previously unseen) data samples.

**Input:** $samplesofdata, X, shape = [n\_samples, \_visible]${*}[r]List of Sensitive Terms

**Output:** $shape = [n\_samples, n\_hidden]${*}[r]Negation Excluded List

**Function** `transform`$(Takesintwoinputs : X, anddetails)$**:**

  $X \longleftarrow self.preprocess(X)$

  $p\_y\_given\_x, \_, log\_z \longleftarrow self.calculate\_latent(X, self.theta)$

  $labels \longleftarrow self.label(p\_y\_given\_x)$

  **if** *details is true* **then**

    **return** $returnp\_y\_given\_x, log\_z$

  **end**

  **return** $labels$

**End Function**

**Function** `label`$(p\_y\_given\_x)$**:**

  **return** $(p\_y\_given\_x > 0.5).astype(bool)$

**End Function**



Because it is required that the estimation of the document-topic distributions be obtained as a return value, it is necessary to update the “transform” method, for which its counterpart of the GuidedLDA algorithm was taken as a reference in the “guidedlda.py” module, this method applies topic modeling using Latent Dirichlet Assignment (LDA) on a document term matrix X. In this case, the transform() algorithm 5 takes the document term matrix X as a numpy array and the parameters “max_iter” and “tol” to control the convergence of the model. Stores the topic distribution for each document in the corresponding row of the doc_topic array and returns this array containing the probability values corresponding to the topic distribution for each document.

Once the changes have been made, the module is imported again and generates the results with the probabilities by topic and dominant topic, as can be seen in Table 6.6.

In addition to the probabilities obtained, we label the first 7,113 records with 1 to indicate that these documents are fraud-related and the remaining 7,113 with 0 to indicate otherwise. A filter by dominant topic is applied to this new representation of the dataset 6.6, obtaining four

**Algorithm 5** Transform the data X according to previously fitted model

**Input:** $X : array-like, shape(n\_samples, n\_features)$, $max\_iter : int, optional$, $tol : double, optional$

**Output:** $doc\_topic : array-like, shape(n\_samples, n\_topics)$ {*}[r]Point estimate of the document-topic distributions.

**Function** `transform`$(self, X, max\_iter = 20, tol = 1e-16)$**:**

  **if** *isinstance*$(X, np.ndarray)$ **then**

    $X \longleftarrow np.atleast\_2d(X)$

  **end**

  $doc\_topic \longleftarrow (np.empty)(X.shape[0], self.n\_hidden)$

  $WS, DS \longleftarrow lda.utils.matrix\_to\_list(X)$

  **foreach** $d \in np.unique(DS)$ **do**

    $doc\_topic[d] \longleftarrow self\_transform\_single(WS[DS == d], max\_iter, tol)$

  **end**

  **return** $doc\_topic$

**End Function**



| Docs | Pressure | Opportunity | Rationalization | Others | DT |
|---|---|---|---|---|---|
| Doc 0 | 0.02 | 0.82 | 0.16 | 0.00 | 1 |
| Doc 1 | 0.13 | 0.36 | 0.02 | 0.49 | 3 |
| Doc 2 | 0.25 | 0.65 | 0.10 | 0.01 | 1 |
| Doc 3 | 0.01 | 0.66 | 0.34 | 0.00 | 1 |
| Doc 4 | 0.00 | 0.34 | 0.04 | 0.62 | 3 |
| ... | ... | ... | ... | ... | ... |
| Doc 14225 | 0.25 | 0.07 | 0.00 | 0.68 | 3 |
| Doc 14226 | 0.00 | 0.20 | 0.42 | 0.38 | 2 |
| Doc 14227 | 0.00 | 0.00 | 0.30 | 0.70 | 3 |
| Doc 14228 | 0.00 | 0.06 | 0.00 | 0.94 | 3 |
| Doc 14229 | 0.00 | 0.73 | 0.02 | 0.26 | 1 |

**Table 6.6:** Numerical representation of the distribution of probabilities by topic (pressure, opportunity, rationalization, and others) obtained through CorEx modifying the transform() method. To the 14,229 documents that comprise the corpus, an additional column is added that identifies the dominant topic (DT), representing the highest probability that a document belongs to a specific topic.

datasets per topic (pressure, opportunity, rationalization, and others) that served as input to train classification algorithms.

### 6.5.2 Detection of phrases related to fraud

The second stage of the experimentation consists of applying classification methods to the datasets (pressure, opportunity, rationalization, and others) with the probabilities obtained in the previous stage through the application of the semi-supervised algorithms (GuidedLDA and CorEx). To try to predict behaviors suspected of fraud.

Once the dominant topic filters the original dataset, four datasets are generated, labeled as

| Classification Method's | Predictive Accuracy | | | | Mean |
|---|---|---|---|---|---|
| | T1 | T2 | T3 | T4 | |
| Random Forest: AUC | 0.91 | 0.90 | 0.89 | 0.77 | 0.87 |
| Gradient Boosting: AUC | 0.91 | 0.92 | 0.90 | 0.80 | 0.88 |

**Table 6.7:** Random Forest's and Gradient Boosting's performance in predicting if a document is related to fraud was evaluated using the area under the curve (AUC). T1, T2, T3, and T4 are the corresponding datasets for the four contexts where a subject obtained from CorEx predominates.



fraud and non-fraud for all their records. We build models using these new representations and classification algorithms to predict whether a new document inputted into the model is related to fraud. RF and GB algorithms were applied due to their superior performance, as reported in [4].

#### 6.5.2.1 Classifier performance

In the present work, the ROC curve was used to represent the performance of different machine learning models when classifying documents as related or unrelated to fraud. Several metrics, including recall, accuracy, and precision, can be used to assess the performance of a classification model. One of the main weaknesses of these metrics is that they are susceptible to changes in class distribution. When the ratio of positive to negative occurrences in a test set changes, a model's performance may no longer be optimal or acceptable. However, the ROC curve is independent of the class distribution changes [65], so for this type of analysis, it is a frequently used technique [66] [67]. The ROC curve will not change even if there is a change in the class distribution of a test set. This is because the ROC curve is based on the underlying class conditional distributions from which the data is drawn. It plots a model's true positive rate on the y-axis against its false positive rate on the x-axis. It provides a general measure of model performance, regardless of the various thresholds used. The results can be seen in Figure 6.2 but are also presented in Table 6.7.

Based on these findings, Random Forest and Gradient Boosting perform the best, with a mean area under the curve (AUC) of $0.87$ and $0.88$, respectively. These findings imply that our method to identify fraudulent actions based on topic identification using semi-supervised models would be feasible when developing machine learning models.

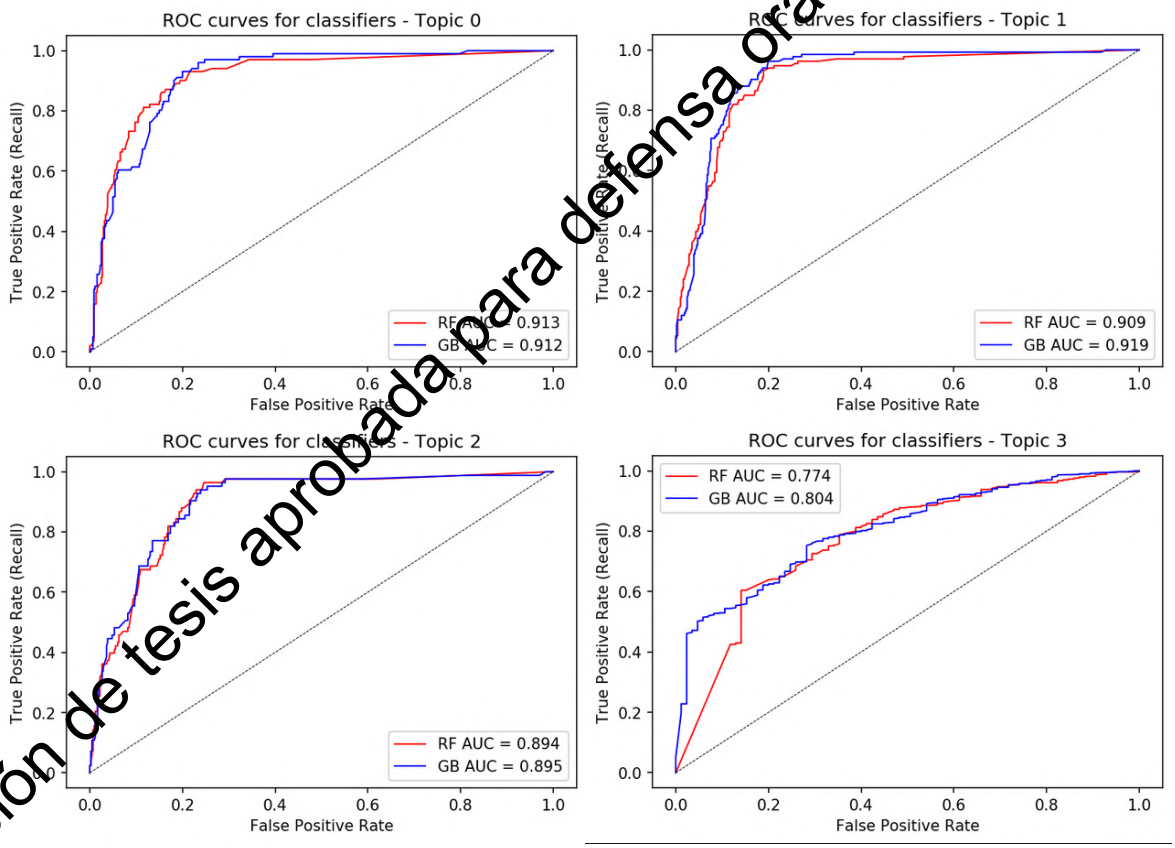




**Figure 6.2:** The ROC curves of different machine learning classification models. The models are: Random Forest (RF) and Gradient Boosting (GB). The results show that GB obtained the highest AUC in all the topics

#### 6.5.2.2 Comparison of classification algorithms.

When comparing the performance of the classification methods, it was observed that RF and GB showed similar performances, with an average AUC of 87 % and 88 %, respectively, as shown in Table 6.7. Furthermore, GB exhibited a slight superiority of 1 % about RF. These findings align with the results reported in [4], where RF and GB were identified as the most efficient classification algorithms, achieving an average AUC of 81 %. By using semi-supervised methods for topic modeling, a notable improvement of 7 % was observed in the performance of the classification methods to predict behaviors suspected of fraud; this suggests that incorporating the semi-supervised approach improves obtaining document probabilities by topic, increasing the accuracy and efficiency of fraud prediction models that use RF and GB algorithms.

#### 6.5.2.3 Validation

The validation of a model consists of evaluating its performance using a dataset that has yet to be used during the training process. The main goal of validation is to estimate a model's performance and get an idea of how well it will work with new data. When building a machine learning model, it is necessary to guarantee its performance through a proper validation process. A standard model validation method uses learning curves and graphs showing the relationship between model performance on training and validation sets as a function of the training data. Observing the relationship between model performance and the amount of training data is possible by analyzing the learning curves. Through cross-validation, it is possible to use k-folds to create a learning curve, train the model on different subsets of data, and evaluate its performance on the validation set. Cross-validation is a technique used to assess the performance of a model by dividing data into k-folds or k-subsets. This allows the model to be trained and evaluated k times, each time on a different subset of data. On the other hand, ROC (receiver operating characteristics) curves provide a way to assess the trade-off between model sensitivity and specificity so they can help determine the optimal threshold for classification tasks. Together, these metrics provide a comprehensive approach to assess and validate the performance of a machine learning model.

Using multiple datasets to validate a model contributes to a more robust estimate of its performance. In this context, four datasets WebScraping, Students, NN, and ChatGPT, were used to perform the tests. Through the application of learning curves, the classifiers (RF and GB) were trained with the four datasets individually. For their validation, the three remaining sets were concatenated, all for each of the four study topics. In other words, four training-validation rounds were carried out, one per dataset; for example, for the first set of tests, the model was trained with WebScraping and validated with (Students+NN+ChatGPT) for each topic, for the three rounds. The remaining datasets were exchanged until all possible combinations were covered. As can be seen in Figure 6.3, a recurring behavior was identified because of applying this technique, observing in the different test rounds that GB has a low bias and acceptable variance in the four topics, which suggests that the model adequately works both in the training set and the test set. Therefore, it can capture the relationship between the characteristics and the objective variable. That is, it does not make assumptions. Furthermore, the model is not sensitive to variations in the training set and can generalize new data well. In the case of RF, it can be mentioned that, in contrast, it has a high bias and

a high variance, which means that the model cannot efficiently capture the relationship between the characteristics and the target variable in the dataset and is sensitive to variations in the dataset so it cannot generalize well to new data.



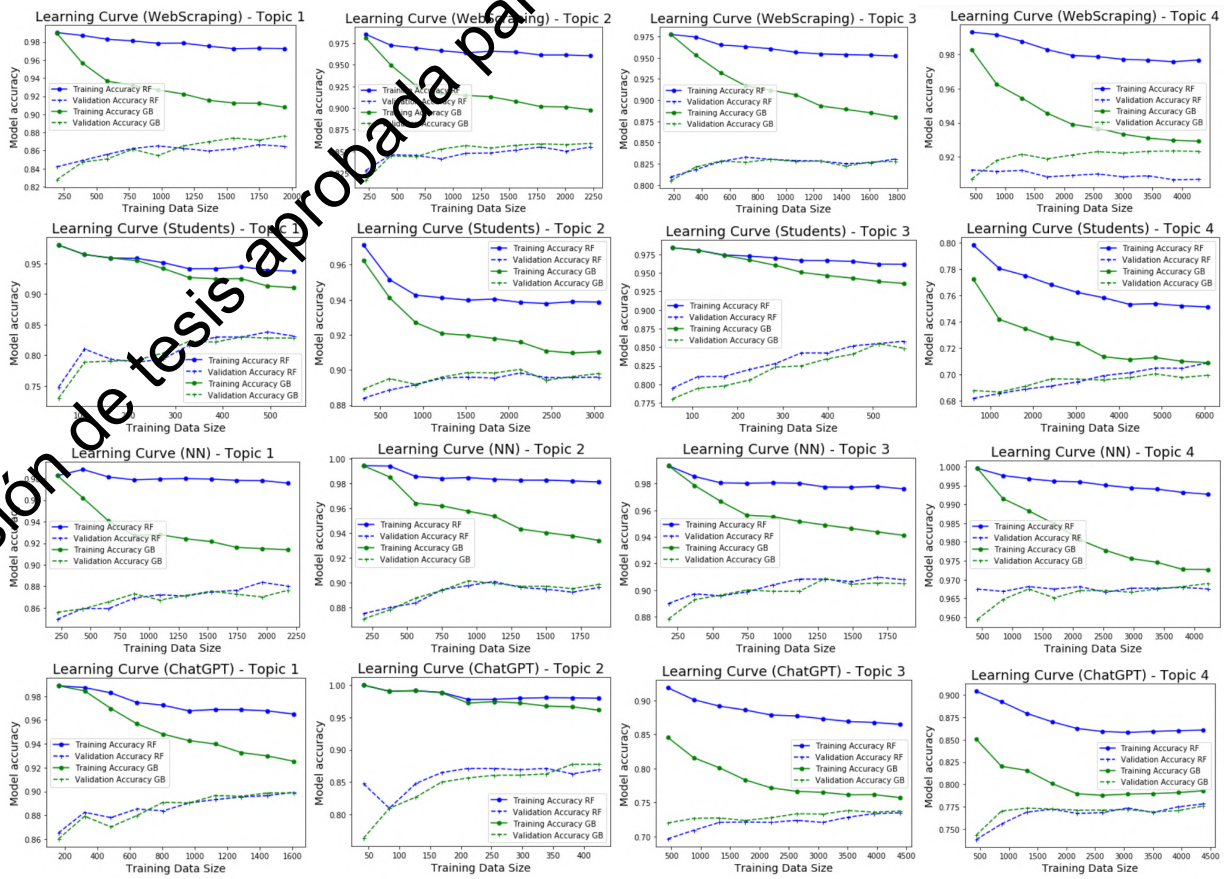


**Figure 6.3:** Learning curves for the four tests were carried out using RF and GB models. This figure also shows the training time of the different models as a function of the size of the training set.

Each of the four datasets was used to train GB once it was verified that its performance was superior to RF, while the remaining three sets were used to test the model's performance. This process was repeated using each of the four training datasets and testing the performance with the remaining three until all possible training-test combinations were covered in the four established topics. ROC curves were generated for each training-test combination to assess the model's performance. By doing this, it was possible to compare the classifier's performance on different datasets and determine which dataset reported superior performance.

| | | Test | | | | |
|---|---|---|---|---|---|---|
| | Dataset | WS | NN | ST | Chat | Dif |
| Train | WS | **0.92** | 0.92 | 0.77 | 0.87 | 0.15 |
| | NN | 0.90 | **0.92** | 0.77 | 0.87 | 0.13 |
| | ST | 0.84 | 0.84 | **0.85** | 0.79 | 0.06 |
| | Chat | 0.90 | 0.82 | 0.83 | **0.93** | 0.11 |

**Table 6.8:** Topic 1

| | | Test | | | | |
|---|---|---|---|---|---|---|
| | Dataset | WS | NN | ST | Chat | Dif |
| Train | WS | **0.93** | 0.92 | 0.89 | 0.83 | 0.10 |
| | NN | 0.90 | **0.89** | 0.86 | 0.80 | 0.10 |
| | ST | 0.87 | 0.86 | **0.93** | 0.89 | 0.07 |
| | Chat | 0.81 | 0.84 | 0.84 | **0.83** | 0.03 |

**Table 6.9:** Topic 2.

| | | Test | | | | |
|---|---|---|---|---|---|---|
| | Dataset | WS | NN | ST | Chat | Dif |
| Train | WS | **0.91** | 0.91 | 0.83 | 0.76 | 0.15 |
| | NN | 0.86 | **0.91** | 0.82 | 0.76 | 0.15 |
| | ST | 0.82 | 0.77 | **0.89** | 0.80 | 0.12 |
| | Chat | 0.83 | 0.80 | 0.80 | **0.82** | 0.03 |

**Table 6.10:** Topic 3

| | | Test | | | | |
|---|---|---|---|---|---|---|
| | Dataset | WS | NN | ST | Chat | Dif |
| Train | WS | 0.82 | **0.82** | 0.66 | 0.68 | 0.16 |
| | NN | 0.79 | **0.82** | 0.65 | 0.69 | 0.17 |
| | ST | 0.60 | 0.60 | **0.78** | 0.70 | 0.18 |
| | Chat | 0.67 | 0.64 | 0.67 | **0.76** | 0.12 |

**Table 6.11:** Topic 4

| | | Test | | | | |
|---|---|---|---|---|---|---|
| | Dataset | WS | NN | ST | Chat | Dif |
| Train | WS | **0.90** | 0.89 | 0.79 | 0.78 | 0.12 |
| | NN | 0.87 | **0.89** | 0.78 | 0.78 | 0.11 |
| | ST | 0.78 | 0.77 | **0.86** | 0.80 | 0.10 |
| | Chat | 0.80 | 0.78 | 0.79 | **0.90** | 0.06 |

**Table 6.12:** Average of the four tests per topic.

| Topics | Folds | | | | | CV |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | Score |
| 1: | 0.85 | 0.85 | 0.84 | 0.84 | 0.86 | **0.85** |
| 2: | 0.86 | 0.84 | 0.85 | 0.85 | 0.85 | **0.85** |
| 3: | 0.79 | 0.78 | 0.77 | 0.78 | 0.77 | **0.78** |
| 4: | 0.80 | 0.79 | 0.79 | 0.79 | 0.80 | **0.79** |

**Table 6.13:** Average Cross Validation (CV) Scores

To identify the behavior the different combinations provide, we can look at the AUC scores obtained for each topic using GB in the four assessments. The higher the AUC score, the better the performance of the classifier. As can be seen in the Tables 6.8, 6.9, 6.10 and

6.11, the combinations that present the best performance are those where the dataset used for training is the same used for testing in all evaluations, obtaining consistently high AUC scores in all the topics, values represented by the main diagonal of each matrix. In those tests where the datasets with which the model was tested differ from those with which the model was trained, we observed that the metrics obtained fluctuate in the combinations made; some have higher AUC values for specific topics and classifiers, while others have lower scores, this suggests that the performance of the classifiers depends on the dataset used for training and testing. Additionally, the values of the four tests were averaged, as seen in Table 6.12, observing the same behavior. In addition, we use the Cross-Validation (CV) technique to contrast the data obtained with the external validation. The dataset was divided into five different "folds," allowing us to train and test the model iteratively. Each iteration used a different fold as a test set, while the remaining folds were used for training. Once all the iterations were completed, the results were averaged to derive a comprehensive performance measure for the model, as seen in Table 6.13. In this context, it is possible to affirm that the model is generalizable since it has been externally validated and cross-validated using the study datasets and the best performance classifier, in addition to the scores in a general way in the different phases of training and test, per topic are consistently high.



#### 6.5.2.4 Discussion

This section compares different topic modeling approaches to capture fraud-related phrases and their computational complexity. The distribution of the main terms and topics obtained from the classic LDA is presented in Table 6.3, with words related to fraud labeled in color. However, fraud-related words are randomly distributed in the topics without any specific clustering, which prevents tagging the topics with the vertices of the FTT. This suggests that the modeling approach cannot determine the relationship between fraud and the FTT. As a result, it cannot be applied on this initial attempt.

Like its unsupervised LDA predecessor, Guided LDA does not show any visible grouping by topic and cannot be associated with the FTT. However, the CorEx algorithm performs highly satisfactorily with grouping words by themes. Table 6.4 shows how words are arranged by a specific color related to the vertex of a fraud triangle, allowing for labeling based on their theme of "pressure, opportunity, and rationalization." This allows a connection between the

FTT and the results obtained by the CorEx model. This suggests that fraud-related phrases within the same individual and corresponding to topics related to the fraud triangle indicate potential fraudsters requiring further investigation.

The dataset is balanced between fraud and non-fraud classes. It is mentioned that analyzing the results with balanced precision or the area under the curve (AUC-ROC) is preferable when dealing with imbalanced data. CorEx shows higher recall, meaning it finds more true positives but has a lower precision or higher false positive rate. Additionally, CorEx outperforms normal and GuidedLDA in terms of balanced accuracy. The semi-supervised approach is considered an alternative strategy to the classic unsupervised model, as it avoids challenges in determining the nature of topics and their labels. Although the topics identified by CorEx do not cover all fraud theories, they align with factors in the FTT, such as “pressure, opportunity, and rationalization.” This approach, incorporating semi-supervised topic modeling techniques and pre-obtained keywords from LDA, is beneficial for identifying relevant topics.

To analyze the computational complexity of LDA, GuidedLDA, and CorEx approaches in discovering latent themes and structures in data, it is essential to understand the practical implications and considerations that researchers should consider when choosing one of these approaches. In the case of LDA, its complexity is influenced by critical factors such as the number of documents (N), the size of the vocabulary (V), and the number of topics (K), with an approximate complexity of O(I * N * K * V) [68], where I represents the number of iterations that an algorithm requires to converge or reach a steady state. The EM (Expectation-Maximization) algorithm used in LDA refers to the number of times the expectation and maximization steps are performed to fit the model to the data. The complexity grows as the number of documents, the size of the vocabulary, and the number of topics increase. This behavior can limit the scalability of LDA on massive data sets or in situations where a high level of thematic granularity is sought. On the other hand, GuidedLDA, by incorporating external information to guide topic assignments, can present additional complexity due to the extra computations required to integrate these guides. However, it follows a similar structure to the LDA in terms of complexity. The benefits of the guide can be remarkable, especially in cases where the interpretability and quality of the topics are a priority. However, an increase in training time can accompany this improvement. In contrast, CorEx differs from other techniques by addressing the correlation and dependency between variables, which impacts its complexity depending on the dimensionality and the number of samples. Since

CorEx operates differently from the probabilistic approach of LDA and GuidedLDA, its complexity is influenced by the number of samples without a fixed number of topic parameters. In summary, the choice between these approaches must consider not only the quality of the results but also the computational complexity and characteristics of the data in question.

## 6.6 CONCLUSIONS

Fraud study and investigation are critical in addressing social disorder and the security threat it poses to government and business. To effectively combat fraud, it is essential to deepen the analysis of fraudulent activities and develop proactive identification strategies. This research used topic modeling and machine learning techniques, focusing on the FTT and using various study corpora. The generation of four datasets was necessary due to the scarcity of resources in this field and the need for fraud-related information. Applying a semi-supervised approach to theme modeling, using the CorEx and GuidedLDA algorithms, demonstrated that CorEx was more successful in creating consistent and interpretable themes aligned with the vertices of the fraud triangle. The probabilities of the document-subject associations extracted from the models were then used as input for the Gradient Boosting and Random Forest classification methods to predict fraud-related behaviors. Evaluation of the model's performance using ROC curves and the AUC metric revealed that Gradient Boosting slightly outperformed Random Forest, achieving an average classification accuracy of 88 % compared to 87 %; This represents a 7 % improvement over the results obtained in a previous study [4]. Semi-supervised approaches like CorEx in text mining contribute to a better analysis of the combination of expertise and domain scalability. Using multiple datasets to test the model's performance yielded promising results, indicating that the model can be generalized. In addition, the model obtained a low bias and an acceptable variance in the four subjects, which indicates good performance in the training and test sets.

### 6.6.1 Future Work

In future work, it is proposed to apply new approaches concerning topic modeling, such as BerTopic, to improve the identification and analysis of relevant topics in large datasets. In this sense, datasets with more information should be generated. This new approach could involve advanced Deep Learning techniques, such as Convolutional Neural Networks or Re-

current Neural Networks, allowing a more precise and contextualized representation of data documents. In addition, incorporating multimodal information, such as images or videos, into topic modeling could be investigated, enriching the understanding of topics by considering different data modalities. In summary, further study and analysis of topic modeling promise innovative approaches that will improve the ability to identify and analyze topics in large volumes of data more accurately.

# REFERENCES

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[2] Roman Egger and Joanne Yu. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology*, 7, 2022.

[3] Shuyang Gao, Rob Brekelmans, Greg Ver Steeg, and Aram Galstyan. Auto-encoding total correlation explanation. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1157–1166. PMLR, 2019.

[4] Marco Sánchez-Aguayo, Luis Urquiza-Aguiar, and José Estrada-Jiménez. Predictive fraud analysis applying the fraud triangle theory through data mining techniques. *Applied Sciences*, 12(7):3382, 2022.

[5] Fauziah Aida Fitri, Muhammad Syukur, and Gita Justisa. Do the fraud triangle components motivate fraud in indonesia? *Australasian Accounting, Business and Finance Journal*, 13(4):63–72, 2019.

[6] Stefania Pecòre. Supporting the annotation experience through corex and word mover's distance. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2021.

[7] P Shamna, VK Govindan, and KA Abdul Nazeer. Content based medical image retrieval using topic and location model. *Journal of biomedical informatics*, 91:103112, 2019.

[8] Beth Lyall-Wilson, Nicolas Kim, and Elizabeth Hohman. Modeling human factors topics in aviation reports. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 63, pages 126–130. SAGE Publications Sage CA: Los Angeles, CA, 2019.

[9] Kyle Reing, David C Kale, Greg Ver Steeg, and Aram Galstyan. Toward interpretable topic discovery via anchored correlation explanation. *arXiv preprint arXiv:1606.07043*, 2016.

[10] HyunSeung Koh and Mark Fienup. Topic modeling as a tool for analyzing library chat transcripts. *Information Technology and Libraries*, 40(3), 2021.

[11] Ryan J Gallagher, Kyle Reing, David Kale, and Greg Ver Steeg. Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics*, 5:529–542, 2017.

[12] Florian Steuber, Mirco Schoenfeld, and Gabi Dreo Rodosek. Topic modeling of short texts using anchor words. In *Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics*, pages 210–219, 2020.

[13] Tarbia Olivier, Iyengar Garud, Bunnell Renée, and Lemaire Alain. Extracting features of entertainment products: A guided latent dirichlet allocation approach informed by the psychology of media consumption. *Journal of Marketing Research*, 56(1):18–36, 2019.

[14] A Hoffmann, M Shi, and U Rüppel. Using topic modeling to restructure the archive system of the german waterways and shipping administration. In *ECPPM 2021–eWork and eBusiness in Architecture, Engineering and Construction*, pages 216–222. CRC Press, 2021.

[15] Cornelia Ferner, Clemens Havas, Elisabeth Birnbacher, Stefan Wegenkittl, and Bernd Resch. Automated seeded latent dirichlet allocation for social media based event detection and mapping. *Information*, 11(8):376, 2020.

[16] Roman Egger and Joanne Yu. Identifying hidden semantic structures in instagram data: a topic modelling comparison. *Tourism Review*, 77(4):1234–1246, 2021.

[17] Marco Sánchez-Aguayo, Luis Urquiza-Aguiar, and José Estrada-Jiménez. Fraud detection using the fraud triangle theory and data mining techniques: a literature review. *Computers*, 10(10):121, 2021.

[18] Rifátul Fitriyah and Santi Novita. Fraud pentagon theory for detecting financial statement fraudulent. *Jurnal Riset Akuntansi Kontemporer*, 13(1):20–25, 2021.

[19] Alcina Augusta De Sena Portugal Dias. Risks and fraud: A theoretical approach. *Revista Perspectiva Empresarial*, 8(2):7–21, 2021.

[20] R Shruti. Exploring the unexplored: A review on forensic fraud. *J Forensic Crime Stu*, 2:103, 2018.

[21] J Moore. Occupational fraud models: A comparative analysis and proposed expanded model. *International Journal of Accounting Research*, 8:203, 2020.

[22] Shaio Yan Huang, Chi-Chen Lin, An-An Chiu, and David C Yen. Fraud detection using fraud triangle risk factors. *Information Systems Frontiers*, 19:1343–1356, 2017.

[23] Novita Puspasari. Fraud theory evolution and its relevance to fraud prevention in the village government in indonesia. *Asia Pacific Fraud Journal*, 1(2):177–188, 2015.

[24] Erna Hidayah and Galih Devi Saptarini. Pentagon fraud analysis in detecting potential financial statement fraud of banking companies in indonesia. *Proceeding Uii-Icabe*, pages 89–102, 2019.

[25] N Christian, YZ Basri, and W Arafah. Analysis of fraud triangle, fraud diamond and fraud pentagon theory to detecting corporate fraud in indonesia. *The International Journal of Business Management and Technology*, 3(4):73–78, 2019.

[26] Ahmad Nurkhin et al. What determinants of academic fraud behavior? from fraud triangle to fraud pentagon perspective. *KnE Social Sciences*, pages 154–167, 2018.

[27] Ike Vayansky and Sathish AP Kumar. A review of topic modeling methods. *Information Systems*, 94:101582, 2020.

[28] Pooja Kherwa and Poonam Bansal. Topic modeling: a comprehensive review. *EAI Endorsed transactions on scalable information systems*, 7(24), 2019.

[29] Olivier Toubia, Garud Iyengar, Renée Bunnell, and Alain Lemaire. Extracting features of entertainment products: A guided latent dirichlet allocation approach informed by the psychology of media consumption. *Journal of Marketing Research*, 56(1):18–36, 2019.

[30] Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 889–892, 2013.

[31] (Guided LDA. (accessed on 8 September 2022).

[32] David Andrzejewski, Xiaojin Zhu, and Mark Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th annual international conference on machine learning*, pages 25–32, 2009.

[33] David Andrzejewski and Xiaojin Zhu. Latent dirichlet allocation with topic-in-set knowledge. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 43–48, 2009.

[34] Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213, 2012.

[35] Sulong Zhou, Pengyu Kan, Qunying Huang, and Janet Silbernagel. A guided latent dirichlet allocation approach to investigate real-time latent topics of twitter data during hurricane laura. *Journal of Information Science*, 49(2):465–479, 2023.

[36] Greg Ver Steeg and Aram Galstyan. Discovering structure in high-dimensional data through correlation explanation. *Advances in Neural Information Processing Systems*, 27, 2014.

[37] Ryan J Gallagher, Kyle Reing, David Kale, and Greg Ver Steeg. Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics*, 5:529–542, 2017.

[38] Jason Sockin. Show me the amenity: Are higher-paying firms better all around? 2022.

[39] Deena Liz John, Ernest Kim, Kunal Kotian, Ker Yu Ong, Tyler White, Luba Gloukhova, Diane Myung-kyung Woodbridge, and Nicholas Ross. Topic modeling to extract information from nutraceutical product reviews. In *2019 16th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, pages 1–6. IEEE, 2019.

[40] Jasmina Dj Novaković, Alempije Veljović, Siniša S Ilić, Željko Papić, and Milica Tomović. Evaluation of classification models in machine learning. *Theory and Applications of Mathematics & Computer Science*, 7(1):39, 2017.

[41] Mohamed Medhat Gaber, Arkady Zaslavsky, and Shonali Krishnaswamy. A survey of classification methods in data streams. *Data Streams: models and algorithms*, pages 39–59, 2007.

[42] Aized Amin Soofi and Arshad Awan. Classification techniques in machine learning: applications and issues. *J. Basic Appl. Sci*, 13:459–465, 2017.

[43] Iqbal H Sarker, ASM Kayes, and Paul Watters. Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. *Journal of Big Data*, 6(1):1–28, 2019.

[44] David J Hand. Assessing the performance of classification methods. *International Statistical Review*, 80(3):400–414, 2012.

[45] Cha Zhang and Yunqian Ma. *Ensemble machine learning: methods and applications*. Springer, 2012.

[46] Gilles Louppe. Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*, 2014.

[47] Jehad Ali, Rehanullah Khan, Nasir Ahmad, and Imran Maqsood. Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, 9(5):272, 2012.

[48] Haowen Deng, Youyou Zhou, Lin Wang, and Cheng Zhang. Ensemble learning for the early prediction of neonatal jaundice with genetic features. *BMC medical informatics and decision making*, 21:1–11, 2021.

[49] Adrián Alcolea and Javier Resano. Fpga accelerator for gradient boosting decision trees. *Electronics*, 10(3):314, 2021.

[50] Zhang Chong, Zhang Xinrui, and Yang Zipei. Enterprise investment value analysis based on machine learning model of rapidminer. In *Journal of Physics: Conference Series*, volume 1584, page 012003. IOP Publishing, 2020.

[51] Wenbao Yu and Taesung Park. Aucpr: An auc-based approach using penalized regression for disease prediction with high-dimensional omics data. *BMC genomics*, 15(10):1–12, 2014.

[52] Tao Wu, Haibin Huang, Guangwei Du, and Yiyong Sun. A novel partial area index of receiver operating characteristic (roc) curve. In *Medical Imaging 2008: Image Perception, Observer Performance, and Technology Assessment*, volume 6917, pages 82–89. SPIE, 2008.

[53] Iain Brown and Christophe Mues. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3):3446–3453, 2012.

[54] Seong Ho Park, Jin Mo Goo, and Chan-Hee Jo. Receiver operating characteristic (roc) curve: practical review for radiologists. *Korean journal of radiology*, 5(1):11–18, 2004.

[55] Jiaqi Guan, Runzhe Li, Sheng Yu, and Xuegong Zhang. A method for generating synthetic electronic medical record text. *IEEE/ACM transactions on computational biology and bioinformatics*, 18(1):173–182, 2019.

[56] Randomwordgenerator. (accessed on 8 September 2022).

[57] Marco Sánchez and Luis Urquiza-Aguiar. Comparative analysis of the performance of machine learning techniques applied to real and synthetic fraud-oriented datasets. In *Doctoral Symposium on Information and Communication Technologies: Second Doctoral Symposium, DSICT 2022, Manta, Ecuador, October 12–14, 2022, Proceedings*, pages 44–56. Springer, 2022.

[58] Christopher Grant Kirwan and Fu Zhiyong. *Smart cities and artificial intelligence: convergent systems for planning, design, and operations*. Elsevier, 2020.

[59] Mohamed Osman Hegazi, Yasser Al-Dossari, Abdullah Al-Yahy, Abdulaziz Al-Sumari, and Anwer Hilal. Preprocessing arabic text on social media. *Heliyon*, 7(2):e06191, 2021.

[60] Ammar Ismael Kadhim. An evaluation of preprocessing techniques for text classification. *International Journal of Computer Science and Information Security (IJCSIS)*, 16(6):22–32, 2018.

[61] Ruba Alnusyan, Ruba Almotairi, Sarah Almufadhi, Amal A Shargabi, and Jowharah Alshobaili. A semi-supervised approach for user reviews topic modeling and classification. In *2020 International Conference on Computing and Information Technology (ICCIT-1441)*, pages 1–5. IEEE, 2020.

[62] Roman Egger and Joanne Yu. Identifying hidden semantic structures in instagram data: a topic modelling comparison. *Tourism Review*, 77(4):1234–1246, 2021.

[63] Greg Ver Steeg. Unsupervised learning via total correlation explanation. *arXiv preprint arXiv:1706.08984*, 2017.

[64] Sebastiaan Merino and Martin Atzmueller. Multimodal behavioral mobility pattern mining and analysis using topic modeling on gps data. In *Behavioral Analytics in Social and Ubiquitous Environments: 6th International Workshop on Mining Ubiquitous and Social Environments, MUSE 2015, Porto, Portugal, September 7, 2015; 6th International Workshop on Modeling Social Media, MSM 2015, Florence, Italy, May 19, 2015; 7th International Workshop on Modeling Social Media, MSM 2016, Montreal, QC, Canada, April 12, 2016; Revised Selected Papers 6*, pages 68–88. Springer, 2019.

[65] Shaomin Wu, Peter Flach, and Cesar Ferri. An improved model selection heuristic for auc. In *Machine Learning: ECML 2007: 18th European Conference on Machine Learning, Warsaw, Poland, September 17-21, 2007. Proceedings 18*, pages 478–489. Springer, 2007.

[66] OP Trifonova, PG Lokhov, and AI Archakov. Metabolic profiling of human blood. *Biomeditsinskaia Khimiia*, 60(3):281–294, 2014.

[67] Susan Mallett, Steve Halligan, Gary S Collins, and Doug G Altman. Exploration of analysis methods for diagnostic imaging tests: problems with roc auc and confidence scores in ct colonography. *PloS one*, 9(10):e107633, 2014.

[68] Koffi Eddy Ihou and Nizar Bouguila. Variational-based latent generalized dirichlet allocation model in the collapsed space and applications. *Neurocomputing*, 332:372–395, 2019.

# 7 DISCUSSION

This chapter contrasts how the work carried out supports the solution to the problem posed in this investigation; this is evidenced by analyzing the results obtained in each publication and how they contributed to meeting the thesisóbjectives. Fraud-related investigations generally focus their attention on figures and financial analysis, dismissing information of a textual nature. In this context, a favorable scenario opens up the study of this phenomenon from the perspective of text analysis that deals with the study of unstructured data.

Schemes and techniques for committing fraud are constantly evolving and are statistically ahead of mitigation strategies. In this sense, the research developed in this thesis proposes applying text mining tools, topic modeling, and classification methods, intending to implement a methodology that allows the early identification of behaviors with suspicion of fraud. To overcome the lack of information in which evidence related to fraud is found, either due to its reserved nature and difficult access or its non-existence, it was necessary to build a synthetically generated dataset, which includes information related to the three vertices of the fraud triangle theory, which are related to fraudulent trends and also included another dimension that refers to general trends. On this basis, in the experimentation stage through the application of topic modeling algorithms, it was possible to identify fraud-oriented behaviors in the study dataset, categorizing the most representative words into four topics, without this meaning that each topic generated by the model is associated with a specific vertex of the triangle of fraud and rather a transversal ordering was observed throughout the topics. The obtained topic model allowed us to deal with probabilities to identify the belonging of a document to a specific topic and to be able to apply classification algorithms to these files to try to predict suspicions of fraud.

## 7.1 CONTRIBUTIONS

Understanding that the problem for the analysis and study of fraud lies in the lack of methods that allow us to build adequate models for its identification quickly and proactively, several interrelated works were developed to propose a solution to this problem. The contributions are described below:

1. Journal-MDPI-1: The preparation of the paper entitled Fraud Detection Using the Fraud Triangle Theory and Data Mining Techniques: A Literature Review published in Computers magazine, Special Edition Artificial Intelligence for Digital Humanities MDPI, 2021. (SJR-Q2). It allowed the development of an SLR, through which it was possible to analyze the different contributions made about fraud in the scientific field and identify the literature related to its detection.

2. Journal-MDPI-2: Once the literature analysis has been carried out, the knowledge gaps and the available scientific evidence on the subject of study have been identified. These results allowed us to understand the problems related to fraud and to distinguish the different theories and methodologies applied to fight it, which allowed us to propose a model for its prediction, trying to identify suspicious behaviors in a dataset and align them with some theory related to fraud. These results were published in the article Predictive Fraud Analysis Applying the Fraud Triangle Theory through Data Mining Techniques in the journal Applied Sciences (JCR-Q2).

3. Journal-IJASEIT: Through the elaboration and results obtained by proposing a model to predict fraud, it was possible to identify the lack of access and availability of information related to this phenomenon, representing a restriction limiting its study. An alternative to obtaining this information is to generate synthetic data, following adequate methodologies that allow for establishing control parameters for a reliable output with all the characteristics of an original dataset. The results were published in the International journal Advanced Science Engineering Information Technology in the article Generation of a Synthetic Dataset for the Study of Fraud through Deep Learning Techniques. (SJR-Q3)

4. Conference-DSICT: Since three datasets (2 synthetics and one real) were generated using different alternatives, it was necessary to compare to establish if this diversity

impacts the results obtained by the prediction model, the different datasets were entered into the model to identify if the performance obtained using synthetic data is comparable to real data and to analyze the behaviors'similarities. This work, named Comparative Analysis of the Performance of Machine Learning Techniques Applied to Real and Synthetic Fraud-Oriented Datasets, was presented at the Doctoral Symposium on Information and Communication Technologies (DSICT 2022) conference.

5. Journal-PeerJ: Once it has been verified that the results of the fraud suspicion prediction model are not distorted by the use of different sets of generated data (real and synthetic) based on the established model, it is proposed to analyze alternatives to improve the performance of the topic modeling considering semi-supervised techniques and using the approach of the previous contribution to validating its impact on the performance of the predictions obtained. The results were submitted for possible publication in the journal PeerJ Computer Science (JCR-Q2)

6. Conference-IEEE: Finally, as an appendix of this thesis, it was proposed to implement a functional architecture model to deploy a solution through the use of open-source tools that allow listening to the information traffic of a network to collect information to carry out the analysis of that data later to try to detect unusual behaviors related to fraud. This proposal was presented at the IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC 2018).



## 7.2 RESEARCH QUESTIONS ANALYSIS

To analyze how the contributions presented in this thesis help solve the problems raised, a discussion is carried out to establish the relationship between the research questions and the contributions mentioned.

RQ1. *What are the advances in fraud detection using topic modeling and machine learning techniques, and how have they been applied to various fraud theories in recent literature?*

Literature reviews are a preliminary step before beginning an investigation since it is a necessary phase that allows us to locate and support the investigation based on what other researchers have written on a specific topic. When a topic needs to be better defined, it is

advisable to make a first foray into the literature to identify what has been written on the topic of study.

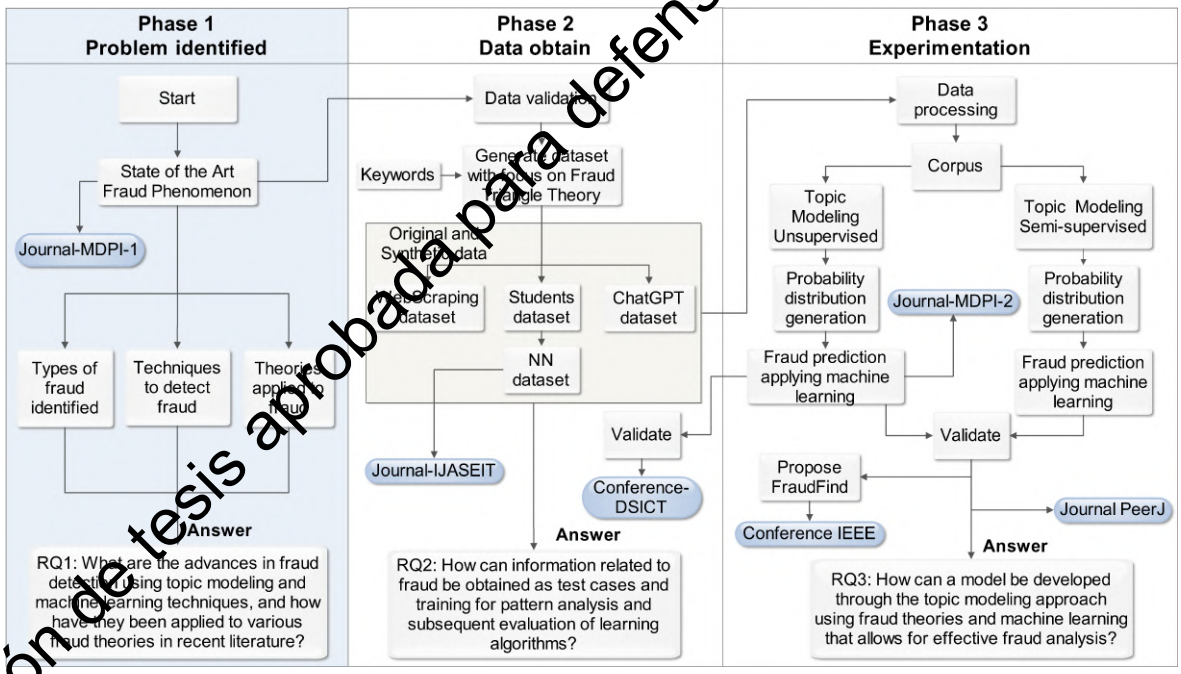


**Figure 7.1:** Contributions - Phase 1.



A systematic review of the literature was carried out in chapter two (Contribution 1, “Journal MDPI-1”, see Phase 1 in Figure 7.1), in which multiple studies related to fraud were collected and analyzed through a systematic process, considering human behavior as the main element of risk. In this context, the theories associated with fraud have reviewed this phenomenon. An explicit contribution of an SLR is that it allows us to obtain a synthesis of information regarding one or several specific research questions under a defined search strategy whose main objective is to find the most significant amount of relevant bibliography available. In this sense, the following research questions were posed:

- RQ1.1: How can fraud be detected by analyzing human behavior by applying fraud theories?
- RQ1.2: What machine or deep learning techniques are used to detect fraud?
- RQ1.3: Using machine learning techniques, how can fraud cases be detected by analyzing human behavior associated with the Fraud Triangle Theory?

Table 7.1 shows the research questions posed and the number of works found with their respective study identifier.

**Table 7.1:** Data extraction form.

| RQ | Study Identifier | Frequency |
|---|---|---|
| 1 | [38, 39, 40, 41, 42, 43, 70] | 7 |
| 2 | [45, 46, 47, 48, 12, 50, 51, 52, 53, 54, 55] [56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68] | 24 |
| 3 | [69] | 1 |



In RQ1.1, seven works were identified based on fraud theories, such as the triangle and the diamond, and tried to identify behavior patterns related to the vertices of these theories. In [38] [39], a model for fraud detection considering the vertices of the fraud triangle is proposed, associating this theory with patterns of human behavior, which can be found in information sources such as emails, text messages, traffic networks, and system logs from which evidence of fraud can be extracted. At the same time, other authors go further [42][70] and analyze the moral aspects of criminology and psychology that can intervene to commit fraud. On the other hand, there are works [41][43] that are based on the use of international standards and regulations related to auditing, which focus on auditors'responsibility when evaluating fraud and whether the standards have been used efficiently based on indicators obtained from surveys carried out on specialized personnel such as auditors, accountants, public officials, and inspectors to determine the perceptions of the importance of the existence of warning signs of financial fraud through the use of fraud theories. In [40], the need to carry out an expost analysis and the existing literature on fraud is identified once information on the behavior of fraudsters has been obtained.

RQ1.2 was the question with the most significant number of works found. A total of 24 were identified, which can be addressed in more detail in Chapter 3. RQ3 was the question with the fewest works found, with only one study linking the detection of fraud through the use of machine learning techniques and that is related to a theory of fraud [69]; in this work, the authors review the aspects related to the vertices of the fraud triangle through the use of data mining techniques to evaluate the attributes such as pressure, opportunity, and rationalization through the use of questionnaires prepared by experts and compared if the suggestions of these agreed with the results obtained through the use of machine learning techniques. Evidence from a single primary paper supporting this approach means a gap in this field. It allows us to establish a baseline for addressing this phenomenon. When identifying the area of investigation related to fraud within the RQ1.3 approach is incipient,

questions were formulated in this area oriented towards more specific areas of knowledge and, above all, establishing the need to build a model.

RQ2. *How can information related to fraud be obtained as test cases and training for pattern analysis and subsequent evaluation of learning algorithms?*

The development of this thesis presents a significant challenge due to the lack of accessibility and scarcity of data related to fraud. Recognizing this limitation, our work contributes to existing knowledge by addressing the difficulty of obtaining datasets that contain the necessary evidence related to this issue. This resource is crucial to evaluate the effectiveness of the proposed model (Contributions 3 and 4, "Journal-IJASEIT and Conference-DSICT," see Phase 2 in Figure 7.2).

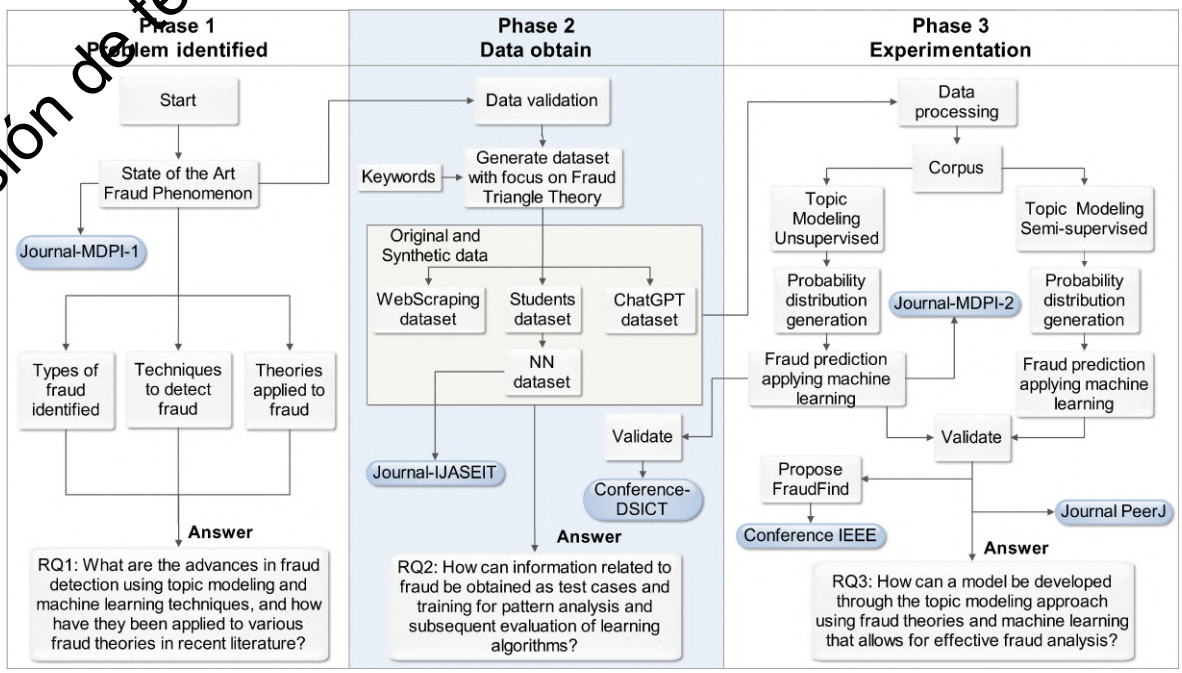


**Figure 7.2:** Contributions - Phase 2.

In Chapter 3, an initial model for fraud prediction is proposed. A synthetic dataset was generated to verify its performance, created from a dictionary of terms related to the fraud triangle. The same triangle is tagged in pressure, opportunity, and rationalization categories. Through the application of online tools, a group of keywords was used, the most representative, and it was used to generate phrases related to the three vertices of the fraud triangle; as a result, the generated phrases include the terms used from the dictionary. Additionally, this process was used to generate phrases unrelated to fraud in the same proportion as those related to fraud and to create a balanced dataset containing phrases in both senses. As a result of

using this dataset in the proposed model, it was possible to observe an acceptable performance in predicting fraud-related behaviors from the perspective of traditional classification models. In contrast, from the deep learning side, the performance was less efficient; this suggests that for this approach, the performance of our model is related to the amount of data and, as such, its size conditions this behavior.

To validate our model under other conditions, chapter 4 proposes generating a new synthetic dataset from a real one. The construction of the real dataset was carried out with the support of EPN students who, depending on the keywords related to the vertices of the fraud triangle, wrote sentences that contained these terms, assuming the role of a potential fraudster, which consisted of asking them if they were in difficult financial conditions, which motivates pressure. They had the opportunity to commit an illegal act on the condition that no one would notice, trying to rationalize the act. They occupy that character and express it in sentences within that context. With this experiment, the real dataset that we named Students was synthesized, the same one that served as a seed to generate a synthetic dataset using a recurrent neural network (RNN) and short-term memory networks (LSTM).

Finally, in Chapter 6, a dataset called ChatGPT was created to simulate features of previous datasets using this artificial intelligence. Phrases related to the three vertices of the fraud triangle were generated: opportunity, rationalization, and pressure. The tool was asked to come up with scenarios that incorporated these elements, such as an employee having access to confidential financial information (opportunity), justifying fraudulent actions due to unfair compensation (rationalization), or facing financial difficulties that motivated fraud (pressure). The tool was then prompted for sentences related to these aspects, posing a question: “Tell me sentences about perceived pressure, opportunity, or rationalization.”

Four different datasets were used to develop and validate a model for fraud prediction. In table 7.2, Let’s explain the information and differences between these datasets:

RQ3. *How can a model be developed through the topic modeling approach using fraud theories and machine learning that allows for effective fraud analysis?*

In Chapter 3, a new model based on topic modeling techniques was proposed that, by applying classification algorithms, tries to predict fraud (Contributions 2 and 5, “Journal-MDPI-2 and Journal-PeerJ”, see Phase 3 in Figure 7.3). As a contribution, the fraud triangle theory was adapted to the proposed model, proposed by the American criminologist

**Table 7.2:** Summary of Datasets Used in Research

| Dataset | Purpose | Key Characteristics | Key Differences |
|---|---|---|---|
| Synthetic Dataset for Initial Model (Chapter 3, “WebScraping”) | Evaluate the initial fraud prediction model. | - Categorized fraud triangle terms - Generated phrases with keywords - Included fraud and non-fraud samples | The dataset was entirely synthetic, generated from keywords and phrases in a dictionary. |
| Real Dataset Generated by Students (Chapter 4, “Students”) | Generate real data for model validation. | - Roleplay scenarios by students - Real human input - Basis for further dataset generation | The dataset was based on real human input, with students assuming roles related to fraud. |
| Synthetic Dataset Generated from Real Dataset (Chapter 4, “Neural Networks”) | Enhance the dataset derived from real data. | - Created with RNN and LSTM neural networks | The dataset was generated from the “Students” dataset using neural networks. |
| Synthetic Dataset Generated by Artificial Intelligence “ChatGPT” (Chapter 6, “ChatGPT”) | Evaluate the fraud prediction model. | - Created with RNN and LSTM neural networks | The dataset was generated from the “Students” dataset using neural networks. |



Donald Cressey in the 1960s, a time since it has not developed substantially, by defining three fundamental factors: pressure, opportunity, and rationalization. This theory has tried to demonstrate how the union of these three factors makes a person commit fraud, which makes it pertinent that the study of fraud is given from the perspective and vision of the fraudster. For this, it is almost indisputable that the analysis of human behavior plays an essential role in detecting potential fraudsters. This will allow us to assign labels related to the vertices or factors of the fraud theory in the dataset. This novel approach provides a methodology to capture the behavior of a potential fraudster using feature classification in the context of the dataset to try to capture their behavior using topic modeling, specifically unsupervised methods such as LDA. Incorporating this technique associated with the fraud triangle theory will allow identifying fraud-related patterns in the study corpus, trying to establish a relationship between the model obtained by LDA and the “pressure, opportunity, and rationalization” vertices. This first approximation could not be established because, in the distribution of words by resulting topics, a dispersion of the words related to fraud in the different topics was observed, which did not allow relating the topics obtained with a specific vertex of the triangle. However, this accumulation of prevalent terms in the different topics

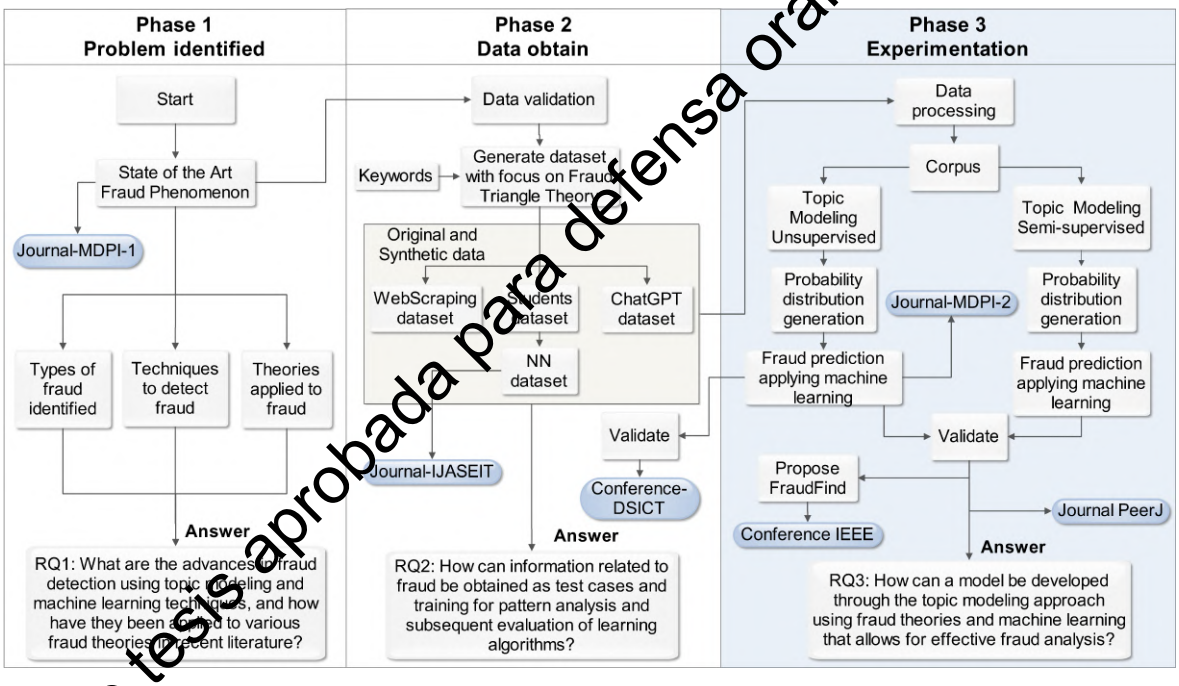


**Figure 7.3:** Contributions - Phase 3.



warns of a visibly fraud-oriented behavior, which suggests that the documents used by the model have a clear probability of fraudulent behavior.

To ensure that the results obtained by the model are not biased because the datasets used for this type of study, such as fraud analysis, are frequently unbalanced. Considering that this data is used in the training and testing phases. It was decided to balance the data in a 1:1 ratio, which means 50 % fraud data and 50 % non-fraud data, which allows us to solve the problem of unbalanced class distribution. In addition, the dataset for testing and training later used in classification methods and deep learning algorithms was split 80:20 respectively. The experimentation shows a promising prediction result, for which the ROC-AUC curves were used as a metric that allows measuring the performance of classification algorithms, obtaining a value of 81 % for the algorithms that presented the best performance.

Chapter 6 presents the scope of the previously proposed method, in which semi-supervised models are proposed to improve efficiency in identifying patterns of ordinary meaning within a document in the modeling of topics stage. This approach aims to integrate the knowledge of the study domain through keywords or anchor words, in our case related to fraud, which guides or encourages modeling in the direction of those words, which means that the algorithm tries to search for topics related to these anchor words, this to find topics of interest related to the three vertices of the fraud triangle “pressure, opportunity and rationalization” and an additional factor named as others if no document is linked to said vertices. As a

contribution in this experimentation phase, it was possible to observe that the distributions of words per topic obtained were more consistent, evidencing in each grouping set of terms related to the vertices of the fraud triangle. Therefore, in this second instance, a relationship can be established between the model obtained by the semi-supervised algorithm (CorEx) and the vertices of the fraud triangle, allowing us to assign the corresponding labels to each topic. With the results obtained by the model based on the probabilities that a document belongs to a specific topic and applying classification methods on these distributions, a much better prediction result was identified than its predecessor, obtaining a performance of 90 % in the algorithms with better performance.

The validity of the work presented in this thesis poses several challenges. Therefore, it is essential to analyze the factors that threaten its reproducibility and reliability to mitigate them. For this purpose, in [1], the datasets and the analysis notebooks with the results in the different research phases are provided.

As a first validation scenario, we propose to analyze the datasets generated for this research to evaluate how they behave based on their performance once the proposed fraud prediction model has been used. Chapter 5 deals with this comparison, establishing the need to identify whether using different datasets can impact the performance of the proposed model. Two datasets used for this study were synthetically generated using neural networks and tools available on the internet. At the same time, a third was developed with EPN students, called students, due to the conditions in which it was built. These resources served as input for the fraud prediction model, which allowed evaluation of them and determination that the performance of the different models generated by each dataset (synthetic and real) was similar. This suggests that the results obtained by synthetic datasets may reflect behaviors as if real data had been used.

Internal validation of a model involves using a subset of the training data to evaluate the model's performance. This is typically done using techniques such as k-fold cross-validation, where the data is divided into k subsets. The model is trained and evaluated k times, with a different subset of the data used as the validation set each time. External validation of a model involves using a completely independent dataset not used during the internal validation or training process to assess the model's performance. This is important to estimate the model's generalization performance and to know how well the model will perform on unseen data. In this context, in Chapter 6, the proposed model is validated using various datasets. The model was trained by performing training-test combinations with the different datasets

one at a time. Each resulting trained model was used to predict the remaining test sets. The different predictions obtained were compared with the real values, which means that the test dataset corresponded to the training one, with which the model can be evaluated and determined that it generalizes adequately when new data is used. These findings highlight the potential of this machine-learning model to identify fraud. The average AUC from the external validation dataset was slightly less than the original one.

Additionally, Chapter 6 discusses the computational complexity of the three approaches used, LDA, GuidedLDA, and CorEx, to discover latent themes in the data. The complexity of LDA depends on factors such as the number of documents, vocabulary size, and topics, which limits scalability on large data sets. GuidedLDA introduces additional complexity due to integrating external information but offers interpretability benefits. The complexity of CorEx depends on the dimensionality of the data and the sample size, as it operates differently than LDA and GuidedLDA. The choice between these approaches must consider both the quality of the results and the computational complexity.

# REFERENCES

[1] GitHub. (Date last accessed 04-May-2023).

# 8 CONCLUSIONS

This thesis contributes to early fraud detection. In this sense, no studies have addressed this problem from the context of human behavior analysis that uses fraud theories associated with topic modeling and machine learning techniques. This implemented methodology has allowed the identification of patterns of fraudulent behavior related to the fraud triangle and its vertices (pressure, opportunity, and rationalization." Given the absence of fraud-related information, the datasets used for this study were synthetically generated, two of which were created using online tools and machine learning techniques. Moreover, another was generated by simulation with EPN students. Finally, as an additional contribution, a dataset was generated using the ChatGPT tool, with the same characteristics as the previous ones, which represents a fundamental contribution to the development of this research, in contrast to studies for fraud detection related to fraud theories, where the extraction of patterns and the construction of models depend on the static parameterization of their factors or vertices. Therefore, the present investigation analyzes how text-based patterns can vary depending on the nature of the information and, through topic modeling, identify implicit themes in the datasets used. This way, it is possible to classify texts and find relevant patterns related to an object of study, fraud. The texts present a variety of topics, and these topics are expressed through words. This technique informs what topics exist in the collected texts analyzed and what words make them up so that later, a researcher decides what theme is related to each identified topic. This way, the fraud prediction model is automated from the initial data collection to the final results.

## 8.1 THEORETICAL ASPECTS

To test the model's performance, we used several datasets balanced by pairing observations with fraud and non-fraud; these samples were generated in different approaches to bring them closer to the most realistic possible scenario of the probability of fraud. This alternative

for the construction of different data sources incorporating characteristics related to a theory of fraud has made it possible to approach the study of this phenomenon from a psychological perspective and to identify that this particular research design is novel and no related work has been evidenced that follows this methodology. The different approaches for generating the datasets used in the research allowed for obtaining various sources of information, which contributes to strengthening characteristics such as the model's precision, reliability, and generalization.

By using an unsupervised approach to analyze topics to identify latent fraud-related issues in the first instance, LDA was used, obtaining a relatively accurate approximation of the topics and keywords discovered with the different probabilities that a specific corpus document belongs to one of these topics. Additionally, the performance for detecting fraud-related textual patterns was tested using various basic classification techniques such as Naive Bayes and k-neighbors and other more sophisticated ones such as SVM and neural networks. They obtained highly efficient and reliable detection results with an average AUC of 0.81 in the analyzed datasets. With the same approach, in a second instance, semi-supervised models (GuidedLDA and Anchored Corex) were used to characterize the themes in the datasets, understanding that detection performance can be improved by combining unsupervised and semi-supervised topic modeling techniques. They found that the semi-supervised Anchored CorEx approach was more interpretable and produced more coherent themes, leading to significantly superior results. It outperformed LDA on the ROC-AUC evaluation metric, with an average AUC of 0.91.

Due to the different techniques used to generate data, the nature of the information, the focus aimed at trying to detect possible cases of fraud, and the fact that this research is based on fraud theories to locate the different behaviors related to this problem. Other works in this research field focus on the analysis of financial statement fraud. However, from the little existing evidence on similar studies related to the design mentioned above particularities, significant advances can be observed that the results of this research offer. However, from the limited existing evidence on similar studies related to the mentioned design peculiarities, several advances can be observed in this area of study, which shows that this work is a step forward in the field. The findings of this research offer knowledge and contribute to the understanding of the subject.

## 8.2 PRACTICAL ASPECTS

The applicability constitutes one of the essential characteristics of this study since, in the reviewed literature, the related works have empirically tested their models for detecting fraud in samples obtained from any repository, without these proposals having developed a test environment that gets closer to a real scenario.

The findings obtained in this research constitute a contribution to the community related to the study of this problem. In the scientific field, it can improve fraud detection methods, develop adequate data analysis, and contribute to implementing more robust internal controls. On the other hand, it can increase awareness about its constant growth, encouraging researchers to deepen their study of this phenomenon. It is also important to highlight that it can help to develop best practices for detecting and preventing fraud through sharing and disseminating results obtained in related works. It also establishes a link that fosters collaboration between institutions and control bodies. In the business field, it can help identify weaknesses in a company's systems and processes that allow this crime to occur. By understanding how the fraud was perpetrated, the company can implement measures to prevent similar incidents from occurring in the future. If the fraud investigation leads to the identification of the perpetrators, some or all of the losses suffered may be recovered; this may be done through civil litigation or criminal proceedings, depending on the circumstances. In addition, it can contribute so that its detection and prevention occur before it translates into significant losses. Training programs can also be developed that educate employees on identifying and preventing fraud.

Finally, it can be concluded that the model proposed for detecting possible fraud intentions from the point of view of human behavior allows the identification of fraudulent behaviors with high precision and efficiency. The experiment results have shown high performance in detection and increased the understanding of managing classification models in conjunction with topic analysis and fraud theories.

## 8.3 METHODOLOGICAL ASPECTS

The combination of Design Science Research (DSR) and Cross-Industry Standard Process for Data Mining (CRISP-DM) methodologies in the research process for this thesis, in gene-

ral, is very effective in addressing complex and multidisciplinary problems.

Combining both methodologies provided a structured and rigorous approach to the research process. DSR made it possible to develop an innovative solution that met the objectives set. At the same time, the CRISP-DM methodology allowed us to explore and discover patterns in the data, which allowed us to obtain a deeper understanding of the problem. Overall, the combined use of these two methodologies helped to ensure that the research objectives were achieved effectively and efficiently.



## 8.4 FUTURE WORK

Due to the particular limitations evidenced in the present investigation, multiple alternatives for future contributions are mentioned in this section. On the one hand, the literature review could focus on the availability and access to datasets related to fraud. In this context, generating a dataset for the analysis and study of fraud can be further explored. For example, using quality parameters on the developed synthetic data represents an estimate of how well they are generated and whether they maintain the same properties as the original dataset; this will allow the identification of higher scores, which can infer more excellent utility. On the other hand, analyzing the number of original data records used for generation is necessary since they can directly affect the quality of the synthetic data. The more examples are available when training a model for data generation, the easier it is for the model to learn the distributions and correlations in the data accurately. Obtaining a more significant number of example sentences for training is essential, as it would significantly improve completeness in a generation. Additionally, more synthetic records could be generated; this will make it easier to show if the integrity of the data remains intact, since if a parameter such as the synthetic data quality score is low, more synthetic records should be generated, and thus the way it could be deduced if there is a quality problem in the generated data.

In addition, the theoretical concept of fraud could be deepened and other theories evaluated. It should be noted that the most recent and minor applied theories lack practical justification, so most of the related works adopt the fraud triangle as theoretical support. However, this does not prevent other variations from being analyzed since the development of new fraud theories can help to increase the understanding of this phenomenon and offer a more specific vision of some aspects that classical theories generally do not address. It should be considered that not all the vertices of the fraud theories can be evidenced in the available

study data sets, which means it is necessary to identify sources of information that contain these unverified factors that allow this knowledge to be extracted and used to generate new datasets.

Another field with potential for future research is related to classification methods. In this context, some works have applied strategies to improve performance, such as using various classifiers, where individual methods are combined to form a meta-classifier. These assembled methods aim to outperform individual classifiers, also known as base classifiers. The predictions of the base classifiers serve as input for a meta-classifier, and its output will be the final class predicted by the combination of the different classifiers.

# A FRAUDFIND: FINANCIAL FRAUD DETECTION BY ANALYZING HUMAN BEHAVIOR

Marco Sánchez[1*], Jenny Torres[1], Patricio Zambrano[1], Pamela Floreso[1]

[1]Departamento de Informática y Ciencias de la Computación, Escuela Politécnica Nacional, Ladrón de Guevara E11-253, Quito 170517, Ecuador

## A.1 ABSTRACT

Financial fraud is commonly represented by the use of illegal practices where they can intervene from senior managers until payroll employees, becoming a crime punishable by law. There are many techniques developed to analyze, detect, and prevent this behavior being the most important *the fraud triangle theory* associated with the classic financial audit model. To perform this research, a survey of the related works in the existing literature was carried out to establish our own framework. In this context, this paper presents FraudFind. This conceptual framework allows for identifying and outlining a group of people inside a banking organization who commit fraud, supported by the fraud triangle theory. FraudFind works in the approach of continuous audit that will be in charge of collecting information of agents installed in user's equipment. It is based on semantic techniques applied through the collection of phrases typed by the users under study for later being transferred to a repository for later analysis. This proposal encourages to contribute with the field of cybersecurity in the reduction of cases of financial fraud.

**KEY WORDS:** Bank fraud; triangle of fraud; human factor; human behavior

## A.2 INTRODUCTION



Fraud is a worldwide phenomenon that affects public and private organizations, covering a wide variety of illegal practices and acts that involve intentional deception or misrepresentation. According to the Association of Certified Fraud Examiners (ACFE) [1], fraud includes any intentional or deliberate act of depriving another of property or money by cunning, deception, or other unfair acts.

The 2016 PwC Global Economic Crime Survey report describes that more than a third of organizations worldwide have been victims of some kind of economic crime, such as asset misappropriation, bribery, cybercrime, fraud, and money laundering. Approximately 22 % of respondents experienced losses of between one hundred thousand and one million, 14 % suffered losses of more than one million, and 1 % of those surveyed suffered losses of one hundred million dollars. These high loss rates represent a rising trend in costs caused by fraud. In organizations, 56 % of cases are related to internal fraud and 40 % to external; this difference is because any individual related to accounting and financial activities is considered a potential risk factor for fraud [2]. When observing the behavior of people in the scope of business processes, it can be concluded that the human factor is closely linked and related to the fraud triangle theory of Donald R. Cressey [3], where three basic concepts: pressure, opportunity, and rationalization; are needed.

Nowadays, there are different solutions in the commercial field [4], [5] as well as the academic field, where some works in progress have been identified[6], [7] aimed at detecting financial fraud. In both cases, these solutions are focused on using different tools that perform statistical and parametric analysis, as well as behavioral analysis, based on data mining techniques and Big Data. Still, none of them solve the problem of detection financial fraud in real time. FraudFind, unlike other proposals, detects, reports, and stores fraudulent activities in real-time through the periodic analysis of the information generated by users for further analysis and treatment.

This paper presents FraudFind, a conceptual framework that allows detecting and identifying potential criminals who work in the banking field in real-time, based on the Fraud Triangle Theory. For the design of the FraudFind framework, some software components related to information processing were analyzed, among them RabbitMQ, Logstash, and ElasticSearch. In addition, the computerization of the triangle of fraud and the use of semantic techniques

will allow for finding possible bank delinquents with a lower false positive rate.

The rest of the document is structured as follows. Section 2 presents the theoretical framework for the definition of Fraud and the concept of the fraud triangle. Section 3 presents the related works found in the literature. Section 4 details the architecture of the model and the prototype to be implemented in future work. Section 5 continues with the discussion, and section 6 concludes with the conclusions and future work.



## A.3 RELATED WORK

This study aims to design an architecture model adapted to the fraud triangle factors, complemented with the human factor, and analyze suspicious behavior to identify possible cases of fraud for future work to carry out its implementation. In this context, several studies were found in the literature contributing to this topic.

Most of the documents address the issue of financial fraud and the different circumstances surrounding it. Nevertheless, identifying people who might be involved in fraudulent activities is a determining factor. The incursion into the behavioral analysis is quoted to [6], whose authors introduce an automatic text mining process by e-mail to detect different types of message patterns. In [7], a generic architectural model is proposed that supports the factors of the fraud triangle. In addition, it performs the classic quantitative analysis of commercial transactions that are already applied as part of the fraud detection audit. The identification and classification of possible fraud by suspicious individuals is a central element of the internal threat prediction model [8]. A key aspect is to classify individuals by focusing on reducing the internal risk of fraud through a descriptive mining strategy [9].

Besides, the experience of auditors plays an important role in the fight against financial fraud. Some work is proposed, which points to the creation of new frameworks that provide systematic processes to help auditors to discover financial fraud within an organization by analyzing existing information and data mining techniques using their own experience and skills [10]. Accordingly, another proposal creates generic frameworks for the detection of financial fraud FFD to evaluate the different characteristics of FFD algorithms according to a variety of evaluation criteria [11].

New approaches detect atypical values by studying and modifying clustering algorithms, such as K-Means, to improve the performance and accuracy in the detection of unusual

values in a dataset [12]. Capturing unusual patterns related to fraudulent activity involves the analysis of the number of variables that can be examined simultaneously the same as technological advances have increased considerably and can be addressed by the use of more sophisticated neural networks increasing the number of neurons and/or layers at the expense of a higher computational cost [13]. An important factor to mention is how expensive it is to detect potential fraudulent transactions manually. For this reason, the FFD is vital for the prevention of the destructive consequences of financial fraud by making a complete comparison of data mining techniques in order to use the best one [14].

Reviewing the literature, it can be concluded that related work does not cover the anticipated detection of fraud since they perform an analysis after the incident occurred. This paper aims to reduce this gap by conducting an online fraud audit by developing a model that will allow the timely identification of suspicious behavior patterns considering the human factor supported by the fraud triangle theory. This prototype is a tool that will allow individuals to be analyzed inside a corporation to identify possible cases of financial fraud.



## A.4 FRAUD AND THE FRAUD TRIANGLE THEORY

In general, there is not an scientific definition of fraud. Nevertheless, it is considered as a subset of internal threats such as corruption, misappropriation of assets, and fraudulent statements, among others [15]. According to ACFE, fraud is defined as [1] *"the use of one's occupation for personal enrichment through the deliberate misuse or misapplication of the employing organization's resources or assets".* However, due to the scope of this paper, only financial fraud will be considered within a banking environment. In financial fraud, there are two types of fraud: internal and external [16]. Internal fraud encompasses a series of irregularities and illegal acts characterized by the intentional deception of fraudsters, leading to the misappropriation of money and other important resources of the company. In the case of external fraud, this is commonly done in the financial statements, which are falsely presented in reports. Most of the known anomalies are due to the weakness of the internal control mechanisms, and in such situations, the fraudsters commit acts of fraud by exploiting these weaknesses.

The occurrence of fraud is best explained with the help of The Fraud Triangle Theory, illustrated in Figure A.1, proposed by Donald R. Cressey, a leading expert in crime sociology who wrote a series of books on crime prevention. Cressey investigates the reasons behind

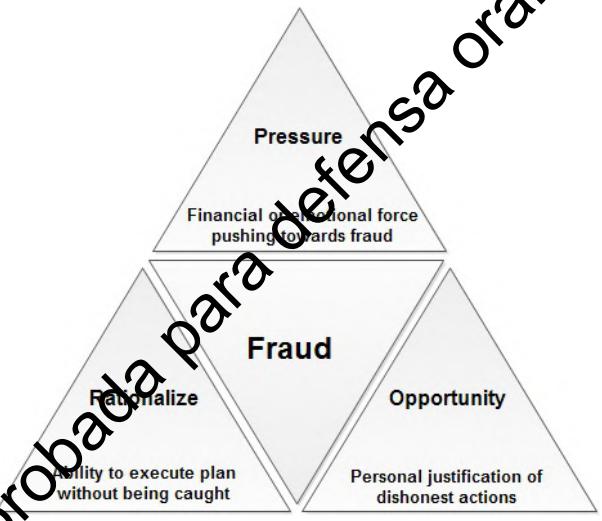


**Figure A.1:** Triangle of Fraud

the question of "why do people commit fraud?" and determines the response in the following three critical elements: perceived pressure, perceived opportunity, and rationalization. Cressey's theory implies that the three elements must be consecutively present to provoke the desire to commit fraud. The first necessary condition in the fraud triangle is the idea of perceived pressure related to the motivation and impulse behind the fraudulent actions of an individual. This motivation often occurs frequently in people under some kind of financial stress. [17]. The second element is the perceived opportunity; and it is the action behind the crime and the ability to commit fraud. Finally, the third component relates to the idea that the individual can rationalize his dishonest actions, making his illegal choices seem justified and acceptable [18]. The risk of committing fraud increases exponentially when there is an increase in the connection between pressure, opportunity, and rationalization.

## A.5 FRAUDFIND FRAMEWORK

The proposed framework operates in the continuous auditing approach to discover financial fraud within an organization belonging to the banking sector, which will be our main study environment. Also, it focuses on the fraud triangle theory, with the human factor considered as an essential element. FraudFind is proposed to analyze large amounts of data from different sources of information for later processing and registration using the ELK stack. ELK is a scalable open-source platform used for real-time data analysis composed by ElasticSearch, Logstash, and Kibana [19] [20] applications, which will be explained below.

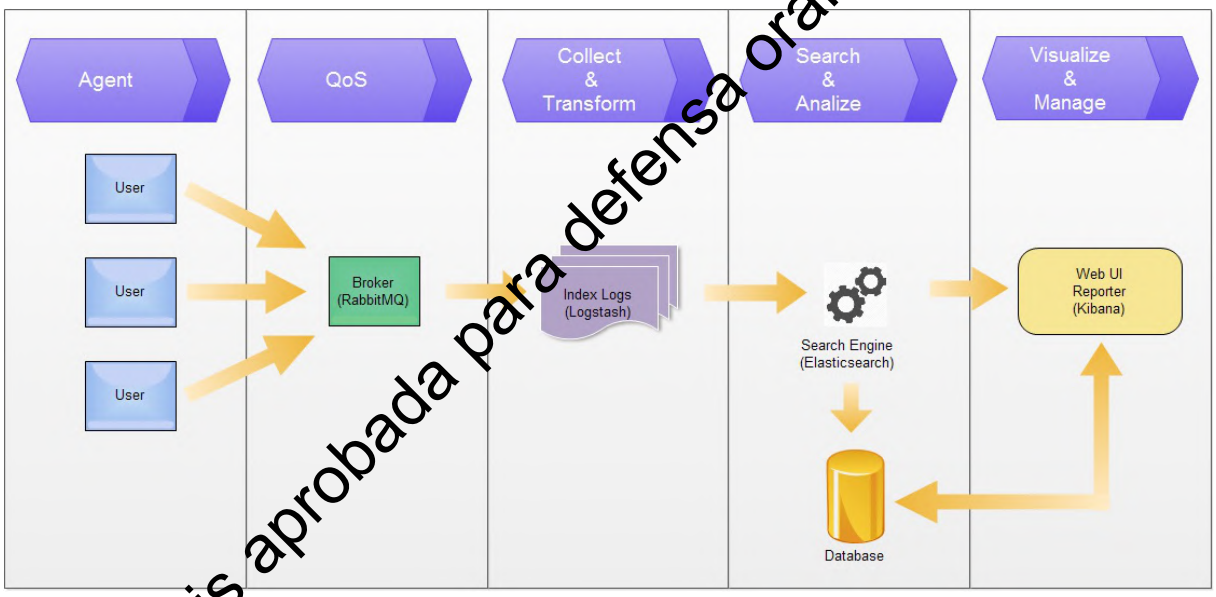


**Figure A.2:** FraudFind Framework



1. ElasticSearch is an open-source search engine developed in Java, which is a distributed, scalable document warehouse and works in real time. Designed mainly to organize data to be easily accessible [21].

2. Logstash is an open-source tool used for event management by centralizing and analyzing many structured and unstructured data types [22].

3. The Kibana web interface is an adjustable board that can be altered and changed to suit our environment. It allows the creation of tables and diagrams, in addition to complex representations [20].

In Figure A.2 we can observe the different modules that compose the framework: Agent, QoS, Collect&Transform, Search&Analyze; and View&Manage.

### A.5.1 Agent

The agent is an application installed in the users'workstations (endpoints) to extract the data they generate from the different sources of information that reside on their equipments. This application is responsible for sending the data entered by the user into RabbitMQ for ordering and classification. Later, this organized information is received by Logstash for its treatment.

### A.5.2 QoS

The integration between several systems or components suggests the need to receive or send information, so these communications must be reliable, safe, fast, and, above all, permanently available. Because the volume of information generated by the agents is considerable and recurrent, this module will ensure its delivery in an orderly and reliable way to Logstash. For this, an intermediary component was introduced, RabbitMQ, to organize and properly distribute the data to for further processing. RabbitMQ is an open-source platform that operates as a message broker, where third-party applications can send and receive messages, offering persistence, confirmation of sending-receiving, and high availability. The cluster of RabbitMQ servers can form a logical broker, allowing the implementation of features such as load balancing and fault tolerance. By default, RabbitMQ sends the messages using the Round - Robin algorithm. After being delivered, it is removed from the queue [23]. Figure A.3 shows the operation of RabbitMQ.

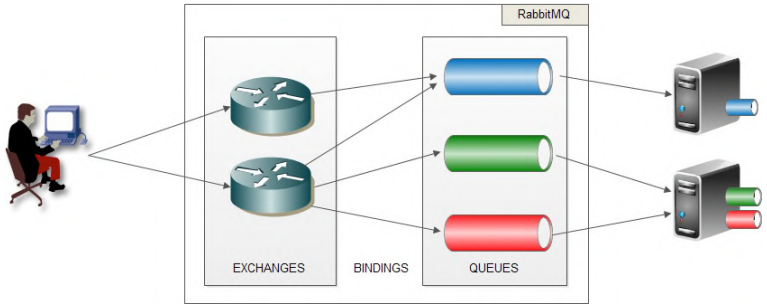


**Figure A.3:** RabbitMQ

### A.5.3 Collect and Transform

This module is responsible for processing the data sent by the agents. As seen in Figure A.2, after ordering the input data of the agents in the QoS module, they are recorded in a temporary file that has raw information that Logstash does not understand and does not know how to handle it. To interpret this information, Logstash has tools called codecs and filters, which perform operations and transformations on the collected data, allowing this information to be converted into a compressible format. Once processed, the information is sent to ElasticSearch for storage. The operation of Logstash is presented in Figure A.4.

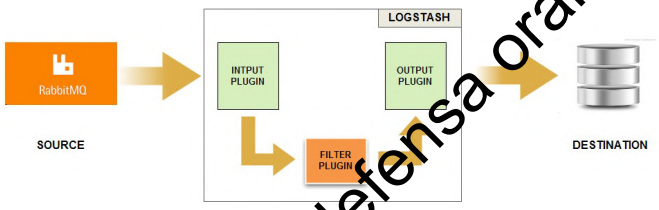


**Figure A.4:** Logstash



### A.5.4 Search and Analyze

This module has all the information processed by Logstash, which is stored immediately after it is received, being able to perform searches efficiently. ElasticSearch is a tool designed with the clustering approach based on the premise of no-fault tolerance hardware. With this property, the information is protected and replicated so that if the physical infrastructure collapses, the data will not be compromised. Figure A.5 shows the architecture of ElasticSearch and its components.

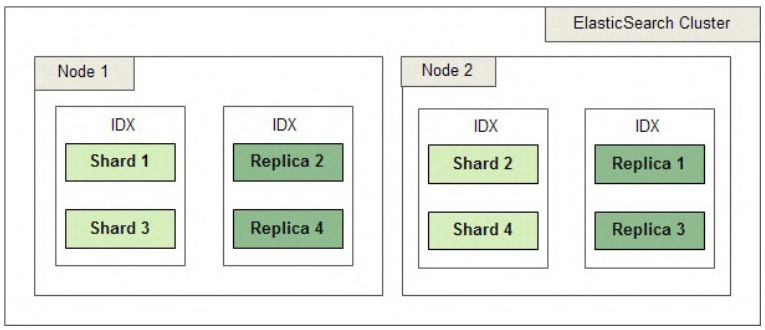


**Figure A.5:** ElasticSearch

### A.5.5 Visualize and Manage

Finally, in this module, the presentation of the data contained in Elasticsearch is performed, using for this purpose Kibana. This tool has been designed to work with ElasticSearch, which allows the visualization and search of information in a customizable way, using histograms, pie charts, and metrics, among others. This tool provides information analysis in real time.

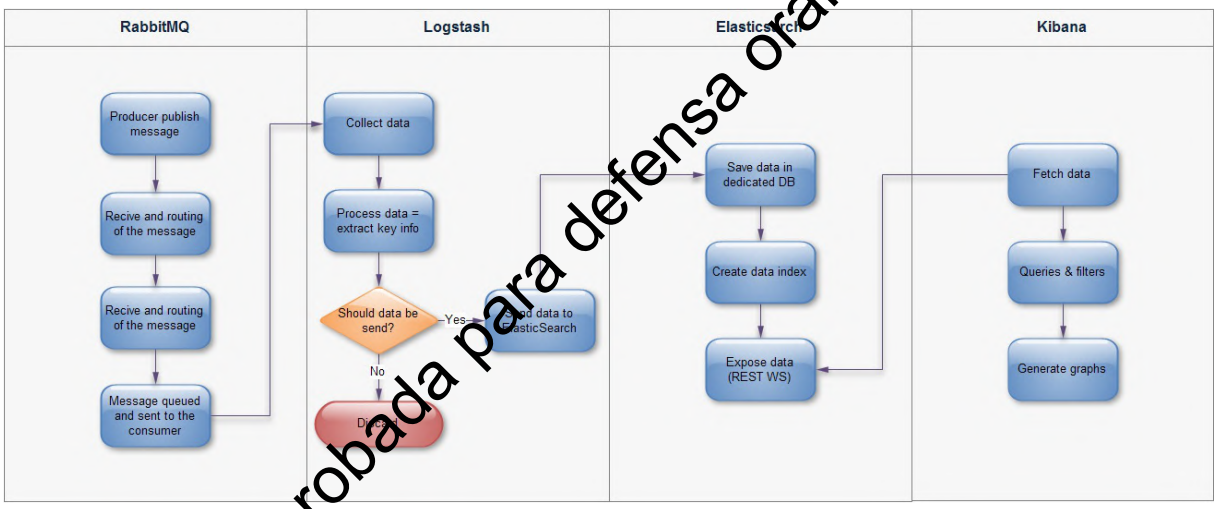


**Figure A.6:** Framework Implementation



## A.6 FRAMEWORK IMPLEMENTATION

In this section, we describe a prototype for the automatic detection of financial fraud, which is currently in the implementation phase. In Figure A.6 we can see the diagram of the proposed framework implementation, which describes the concept of the different modules for practical implementation using free and open-source platforms.

To begin, the information extracted by the agents is sent through data queues, which must be attended to quickly, safely, and reliably. To achieve this goal, RabbitMQ has been used, which is an open-source message broker that implements the Advanced Message Queuing Protocol (AMQP) standard. First of all, several RabbitMQ servers on a local network can be grouped into a logical (distributed) broker. This allows the implementation of features such as load balancing and fault tolerance. Another important feature is the AMQP protocol that RabbitMQ uses, which accepts connections between different platforms.

The data sent by RabbitMQ is received by Logstash for its treatment (organize and categorize). Logstash is a tool that collects, processes, and filters information. According to Figure A.4, it comprises three main plugins: input, filter, and output. First, we have the input plugin that allows the collection of records in different formats, such as files, TCP / UDP, etc. Second, we have the filter plugins that allow Logstash to execute the transformation on the input data. Finally, the output plugin allows processed and transformed data to be written in various formats that go to ElasticSearch [24].

The data sent by Logstash is received by ElasticSearch, which indexes and analyzes this

information. ElasticSearch is a search and storage engine that can handle lots of data in real time, providing speed and reliability [25], along with Kibana as a visualization tool.

Periodically, a task that does the alert tracking checks the information entered and compares it with a fraud triangle library to determine if there is a relation to generating an alert that will be stored in the database. The library of the fraud triangle is just a dictionary that contains three definitions: pressure, opportunity, and justification. Under these parameters, the sentences and words associated with these behaviors are composed.



## A.7 ANALYSIS AND DISCUSSION

### Performance analysis

FraudFind consists of the extraction of data from different sources of information through agents installed in workstations, which collect behavioral data and send this information in an organized way, reporting its activity to the central server. The typed words are sent to RabbitMQ, an application that manages message queues, which delivers fast, secure, and reliable information to Logstash, a tool used to collect and analyze data from monitoring heterogeneous sources, and finally to ElasticSearch, which performs indexing. All this is aimed at ensuring the security of the transactions generated by the users trying to identify possible acts of fraud through the analysis of human behavior and the treatment of the results. Unusual behavior does not guarantee the intentionally of committing fraud, so it should take into consideration the analysis of risk factors associated with this behavior, which should be measurable and weighted in accordance with security policies in an organization.

When there are different sources of information, we find inconsistency in the logs, given that the formats are different. This represents a problem since administrators require access to this information for analysis, and there is difficulty for searching in different formats. When Logs are distributed among the different analysis teams, they are decentralized, and each of them has a different format and different routes to find them, complicating their administration and analysis. ELK solves these problems because it collects all this information to process it, storing it in a distributed manner, and uses treatment techniques such as big data to obtain accurate results.

Additionally, studying human behavior plays an important role in this work. Through this

analysis, it is possible to discover transactions that are part of a pattern not identified in the data traffic, and that would have stopped discovering using traditional means.

### Technical analysis

The ELK (ElasticSearch, Logstash, and Kibana) platform provides versatile and functional records management when searching and analyzing information from a source. Centralized data logging can be useful for identifying unusual traffic patterns, allowing you to search for all stored records that quickly execute the necessary event correlation.

### Security analysis

The possible violation of privacy is a factor that should be considered when implementing this solution within a company. Legal data protection regulations should be considered in a given region. The possible violation of privacy is a factor that must be considered when setting up to integrate this solution into a company. The legal regulations for data protection in a given region should be considered. The level of monitoring will depend on the internal policies in an organization and the laws that are governed in each country and should be determined, taking into account the advice of the legal part of the institution or company.

## A.8 CONCLUSIONS

The present work proposes FraudFind, a conceptual framework to detect financial fraud supported by the fraud triangle factors, which, compared to the classic audit analysis, makes a significant contribution to the early detection of fraud within an organization. Considering human behavior factors, it is possible to detect unusual transactions that would not have been considered using traditional audit methods. These behavior patterns can be found in the information that users generate when using the different applications on a workstation. The collected data is examined using data mining techniques to obtain patterns of suspicious behavior evidencing possible fraudulent behavior. Nevertheless, the legal framework and the different regulations that are applied in public and private institutions of a particular region represent a high risk for the non-implementation of this architecture as an alternative solution. Future work will have as its main objective the implementation and evaluation of

the framework as a tool for continuous auditing within an organization.

# REFERENCES

[1] ACFE Asociacion de Examinadores de Fraudes Certificados. (Date last accessed 15-July-2014).

[2] PwC. (Date last accessed 15-July-2014).

[3] Norman Binti Omar and Hesri Faizal Mohamad Din. Fraud diamond risk indicator: An assessment of its importance and usage. In *2010 International Conference on Science and Social Research (CSSR 2010)*. IEEE, dec 2010.

[4] Lynx. (Date last accessed 15-July-2014).

[5] Ibm. (Date last accessed 15-July-2014).

[6] Carolyn Holton. Identifying disgruntled employee systems fraud risk through text mining: A simple solution for a multi-billion dollar problem. *Decision Support Systems*, 46(4):853–864, mar 2009.

[7] Stefan Hoyer, Halyna Zakhariya, Thorben Sandner, and Michael H. Breitner. Fraud prediction and the human factor: An approach to include human behavior in an automated fraud audit. In *2012 45th Hawaii International Conference on System Sciences*. IEEE, jan 2012.

[8] Miltiadis Kandias, Alexios Mylonas, Nikos Virvilis, Marianthi Theoharidou, and Dimitris Gritzalis. An insider threat prediction model. In *Trust, Privacy and Security in Digital Business*, pages 26–37. Springer Berlin Heidelberg, 2010.

[9] Mieke Jans, Nadine Lybaert, and Koen Vanhoof. Internal fraud risk reduction: Results of a data mining case study. *International Journal of Accounting Information Systems*, 11(1):17–41, mar 2010.

[10] P. K. Panigrahi. A framework for discovering internal financial fraud using analytics. In *2011 International Conference on Communication Systems and Network Technologies*, pages 323–327, June 2011.

[11] D. Yue, X. Wu, Y. Wang, Y. Li, and C. H. Chu. A review of data mining-based financial fraud detection research. In *2007 International Conference on Wireless Communications, Networking and Mobile Computing*, pages 5519–5522, Sept 2007.

[12] M. Ahmed and A. N. Mahmood. A novel approach for outlier detection and clustering improvement. In *2013 IEEE 8th Conference on Industrial Electronics and Applications (ICIEA)*, pages 577–582, June 2013.

[13] A. Vikram, S. Chennuru, H. R. Rao, and S. Upadhyaya. A solution architecture for financial institutions to handle illegal activities: a neural networks approach. In *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*, pages 181–190, Jan 2004.

[14] H. Li and M. L. Wong. Financial fraud detection by using grammar-based multi-objective genetic programming with ensemble learning. In *2015 IEEE Congress on Evolutionary Computation (CEC)*, pages 1113–1120, May 2015.

[15] Dawn Cappelli, Andrew Moore, Randall Trzeciak, and Timothy J Shimeall. Common sense guide to prevention and detection of insider threats 3rd edition–version 3.1. *Published by CERT, Software Engineering Institute, Carnegie Mellon University, http://www. cert. org*, 2009.

[16] Prabin Kumar Panigrahi. A framework for discovering internal financial fraud using analytics. In *2011 International Conference on Communication Systems and Network Technologies*. IEEE, jun 2011.

[17] Grace Mui and Jennifer Mailley. A tale of two triangles: comparing the fraud triangle with criminology's crime triangle. *Accounting Research Journal*, 28(1):45–58, jul 2015.

[18] D. Al-Jumeily, A. Hussain, Á. MacDermott, H. Tawfik, G. Seeckts, and J. Lunn. The development of fraud detection systems for detection of potentially fraudulent applications. In *2015 International Conference on Developments of E-Systems Engineering (DeSE)*, pages 7–13, Dec 2015.

[19] S. GVK and S. R. Dasari. Big spectrum data analysis in dsa enabled lte-a networks: A system architecture. In *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, pages 655–660, Feb 2016.

[20] T. Prakash, M. Kakkar, and K. Patel. Geo-identification of web users through logs using elk stack. In *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, pages 606–610, Jan 2016.

[21] U. Thacker, M. Pandey, and S. S. Rautaray. Performance of elasticsearch in cloud environment with ngram and non-ngram indexing. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pages 3624–3628, March 2016.

[22] Dong Nguyen Doan and Gabriel Iuhasz. Tuning logstash garbage collection for high throughput in a monitoring platform. In *Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2016 18th International Symposium on*, pages 359–365. IEEE, 2016.

[23] V. M. Ionescu. The analysis of the performance of rabbitmq and activemq. In *2015 14th RoEduNet International Conference - Networking in Education and Research (RoEduNet NER)*, pages 132–137, Sept 2015.

[24] D. N. Doan and G. Iuhasz. Tuning logstash garbage collection for high throughput in a monitoring platform. In *2016 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pages 359–365, Sept 2016.

[25] X. M. Li and Y. Y. Wang. Design and implementation of an indexing method based on fields for elasticsearch. In *2015 Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC)*, pages 626–630, Sept 2015.