Graduate Program in Applied Computing

Applied Computing

Academic Master

Luciano Ignaczak

A Value-based Approach for Information Classification

São Leopoldo, 2024

Luciano Ignaczak

A VALUE-BASED APPROACH FOR INFORMATION CLASSIFICATION

Doctorate Thesis presented as a partial requirement to obtain the PhD's degree from the Applied Computing Graduate Program of the Universidade do Vale do Rio dos Sinos — UNISINOS

Advisor:
Prof. Cristiano André da Costa, PhD

São Leopoldo
2024

I24v Ignaczak, Luciano.

A value-based approach for information classification / by Luciano Ignaczak. – 2024.

107 p. : il. ; 30 cm.

Doctorate thesis (PhD’s degree) — Universidade do Vale do Rio dos Sinos, Applied Computing Graduate Program, São Leopoldo, RS, 2024.

“Advisor: Cristiano André da Costa, PhD”.

1. Information classification. 2. Information security. 3. Text mining. 4. Natural language processing. 5. Information value. I. Title.

UDC: 004.89

Cataloging in Publication (CIP):
Librarian Alessandro Dietrich - CRB 10/2338

ACKNOWLEDGEMENTS

I extend my sincerest gratitude to those who have contributed to the completion of this thesis. First and foremost, I express my deepest appreciation to my wife, Gabriela Ignaczak, and my daughter, Lorena Ignaczak. Their unwavering support, understanding, and patience have been invaluable throughout this journey.

I am also profoundly grateful to my parents, João Carlos Ignaczak and Elcida Maria Mezzomo Ignaczak, for instilling in me a love for learning and providing unwavering encouragement throughout my academic pursuits.

Special thanks are due to my advisor, Cristiano André da Costa, whose guidance, expertise, and unwavering support have been instrumental in shaping this research. His insightful advice, encouragement, and commitment to excellence have been invaluable in fostering my research skills and academic growth.

I would also like to extend my gratitude to my esteemed research partners, Marcio Garcia Martins, Felipe Zeizer, Blanda Mello, Guilherme Goldschmidt, Rafael Kunst, and Bruna Donida. Their collaboration and intellectual contributions have enriched this journey immeasurably. Marcio Garcia Martins deserves special mention for his dual role as a research partner and business ally, enhancing both professional and personal dimensions of this endeavor.

I am indebted to UNISINOS and Grupo Hospitalar Conceição for granting me access to invaluable resources and documentation essential for conducting this research.

Lastly, I extend my heartfelt thanks to all my friends for their encouragement, understanding, and unwavering belief in my abilities. Their companionship and camaraderie have enriched my life in countless ways.

ABSTRACT

The digital transformation has revamped how products and services are produced and traded in the digital world. This innovation, bolstered by emerging technologies and evolving business models, underscores the growing importance of information to organizations and amplifies the significance of protecting it. However, a current challenge faced by information security teams is identifying which information requires safeguarding. Today, securing all information collected and produced by an organization is complex due to several limitations, such as budget constraints and understaffed security teams. Furthermore, organizations hold much information that does not require protection. Information classification is the cornerstone process to deal with this challenge in an organization. This process distinguishes confidential from non-confidential information and defines different sensitivity levels. Information classification is a previously introduced research topic, but its real-world application encounters several difficulties due to its manual nature. In order to overcome real-world barriers, scientific research has evaluated the application of natural language processing to automate the process. Most scientific studies proposed supervised learning approaches, which also present drawbacks, such as the significant effort to annotate sensitive labels and the limited flexibility for changes in the information classification scheme. Thus, this study proposes a new information classification model based on the information value. To the best of our knowledge, this is the first attempt to estimate the information value using textual features in the information classification context. The model assesses document value from two perspectives: (i) personal information associated with laws and regulations and (ii) confidential information related to the organizational context. The model applies information extraction and topic modeling to acquire document features and a regression model to estimate information value. We evaluated the proposed model by designing three experiments. The first experiment assessed the performance of two named entity recognition approaches and a relation extraction technique for identifying personal and sensitive personal data. We also implemented an experiment to evaluate the bag-of-words approach to classify documents into four departments. The third experiment assessed the model implementation using a corpus comprising 197 documents from an organization related to the educational sector. The proposed model evaluation implemented six experimental scenarios comprising three, four, and five-level information classification schemes. The model implementation using a Decision Tree regressor achieved an accuracy higher than 80% in the six scenarios. The study also presented that the BERT model outcome LSTM neural network in discovering personal data entities. Finally, the study demonstrated the feasibility of implementing a specific model to extract topics from each organization department since the text classification task achieved an accuracy that did not significantly impact the proposed information classification model.

Keywords: Information Classification. Information Security. Text Mining. Natural Language Processing. Information Value.

RESUMO

A transformação digital está modificando a forma como produtos e serviços são produzidos e negociados no ambiente virtual. Esta mudança tornou as informações mais valiosas às organizações e ampliou a importância de protegê-las. No entanto, um desafio enfrentado pela área de segurança da informação é identificar as informações que necessitam de proteção, pois equipes e orçamentos de segurança da informação possuem limitações. Além disso, organizações possuem informações que não necessitam de proteção. A classificação da informação é o processo responsável por distinguir o nível de sensibilidade de uma informação para uma organização. O processo também é responsável pela atribuição de um rótulo à informação para registrar o nível de sensibilidade. Apesar da classificação da informação não ser um tópico de pesquisa recente, sua aplicação no mundo real enfrenta desafios devido à dependência de pessoas na definição do nível de sensibilidade de uma informação. Para superar estes desafios, pesquisas avaliaram a aplicação de tarefas de processamento de linguagem natural para automatizar o processo de classificação da informação. A principal abordagem analisada nas pesquisas é baseada no uso de aprendizado supervisionado, a qual também enfrenta dificuldades para implementação em uma organização. O esforço para anotação dos dados e a falta de flexibilidade para realizar ajustes no esquema de classificação da informação são dois exemplos de dificuldades enfrentadas. Este estudo propõe um novo modelo para classificação da informação baseado no valor da informação. Embora esta abordagem seja usada em pesquisas relacionadas à segurança da informação, não há conhecimento da sua aplicação para classificar o nível de sensibilidade de um documento. O modelo proposto estima o valor da informação baseado em duas perspectivas: (i) o valor dos dados pessoais, considerando as leis e regulamentações atuais; (ii) o valor da informação baseado no contexto organizacional. O modelo proposto aplica extração de informações e modelagem de tópicos para obter características textuais de um documento e um modelo de regressão para estimar o valor da informação. O modelo proposto foi avaliado a partir do desenho de três experimentos. O primeiro experimento avaliou a performance de duas abordagens para o reconhecimento de entidades mencionadas e uma técnica para extração de relações para a identificação de dados pessoais e dados pessoais sensíveis. O trabalho também implementou um experimento para avaliar a abordagem de sacola de palavras para classificar documentos relacionados com quatro departamentos de uma organização. O terceiro experimento avaliou o modelo com base em seis cenários experimentais compreendendo esquemas para classificação da informação com três, quatro e cinco níveis. A implementação do modelo proposto usando árvores de regressão alcançou uma acurácia superior a 80% em todos os cenários avaliados. O estudo também apresentou que o modelo BERT obteve uma performance superior a uma rede neural LSTM na descoberta de entidades relacionadas com dados pessoais. Por fim, o estudo demonstrou que a implementação de um modelo para extração de tópicos específico para cada setor é viável, pois a tarefa de classificação de texto atingiu uma acurácia que não impacta significativamente no modelo de classificação da informação.

Palavras-chave: Classificação da Informação. Segurança da Informação. Mineração de Texto. Processamento de Linguagem Natural. Valor da Informação.

LIST OF FIGURES

Figure 1 – An overview of the information classification process. 19
Figure 2 – Text mining tasks applied in the cybersecurity domain. 26
Figure 3 – The research protocol to select studies related to information classification. 31
Figure 4 – The proposed model to classify information based on the information value. 43
Figure 5 – An example of taxonomy presenting PII and SPI according to Brazilian law. 44
Figure 6 – An overview of the experiments designed for evaluating the proposed model. 49
Figure 7 – The set of 23 Brazilian-related PII evaluated in this study. 50
Figure 8 – Example of a four-page public announcement provided by the academic secretariat. 53
Figure 9 – The PII discovery process implemented in the experiment. 54
Figure 10 – Format of files resulting from digitized forms before and after applying OCR and plain text conversion. 55
Figure 11 – The activities performed to prepare the model for value-based information classification. 61
Figure 12 – Examples of personal data value estimation considering the most frequent entity for each group. 62
Figure 13 – Examples of features representing the PII presence in documents. 64
Figure 14 – Process performed to evaluate the results produced for the model. 66
Figure 15 – Precision, Recall, and F1-score results for each model considering the set of 80 documents from the health and education sectors. 71
Figure 16 – Precision, Recall, and F1-score results for each model after performing relation extraction considering the set of 80 documents from the health and education sectors. 73
Figure 17 – The Decision Tree confusion matrices for the scenario sets #1 and #2 considering the three-level information classification schemes. 82
Figure 18 – The Gradient Boosting confusion matrices for the scenario sets #1 and #2 considering the three-level information classification schemes. 82
Figure 19 – The Random Forest confusion matrices for the scenario sets #1 and #2 considering the three-level information classification schemes. 83
Figure 20 – The Decision Tree confusion matrices for the scenario sets #1 and #2 considering the four-level information classification schemes. 84
Figure 21 – The Gradient Boosting confusion matrices for the scenario sets #1 and #2 considering the four-level information classification schemes. 84
Figure 22 – The Random Forest confusion matrices for the scenario sets #1 and #2 considering the four-level information classification schemes. 85
Figure 23 – The Decision Tree confusion matrices for the scenario sets #1 and #2 considering the five-level information classification schemes. 85
Figure 24 – The Gradient Boosting confusion matrices for the scenario sets #1 and #2 considering the five-level information classification schemes. 86
Figure 25 – The Random Forest confusion matrices for the scenario sets #1 and #2 considering the five-level information classification schemes. 86

LIST OF TABLES

Table 1 – An example of the three-level information classification scheme from the Whirlpool Corporation. 19
Table 2 – Keywords used to discover relevant studies. 31
Table 3 – Criteria applied to include or exclude studies in the SLR. 32
Table 4 – Summary of the criteria application to select the related works. 33
Table 5 – Characteristics extracted from state-of-the-art studies related to information classification. 34
Table 6 – Characteristics extracted from state-of-the-art studies associated with PII discovery. 38
Table 7 – An example of a three-level information classification scheme to assign the sensitivity label. 47
Table 8 – The intensities and its descriptions presented to employee for document association. 63
Table 9 – Number of groups and documents categorized into each intensity. 63
Table 10 – The six experimental scenarios defined to assign the sensitivity label to a document. We evaluated two scenarios for each classification scheme considering distinct ranges of values (values in US dollars). 68
Table 11 – Table presents the number of correct, incorrect, partially correct, missed, and spurious entities for each of the four evaluation types. This table considers all entities identified in the documents, even those unrelated to personal data. 70
Table 12 – Precision, Recall, and F1-score results for each model considering the set of 80 documents from the health and education sectors. The best F1 scores are highlighted in bold. 70
Table 13 – Table presents the number of correct, incorrect, partially correct, missed, and spurious entities for each of the four evaluation types after performing entity relation. This table considers only entities related to personal data. . 72
Table 14 – Precision, Recall, and F1-score for each model after performing relation extraction rules considering the 80 documents from the health and education sectors. The best F1 scores are highlighted in bold. 74
Table 15 – Precision, Recall, and F1-score for each model after performing relation extraction rules considering 40 documents from the education sector. The best F1 scores are highlighted in bold. 74
Table 16 – Precision, Recall, and F1-score for each model after performing relation extraction rules considering 40 documents from the health sector. The best F1 scores are highlighted in bold. 75
Table 17 – The number of documents containing PII that each model would not discover, considering the entities identified by the NERC models and after performing the RE approach. 75
Table 18 – Results for the mean and its 95% confidence interval for each preprocessing combination employing the Random Forest model. 78
Table 19 – Results for the mean and its 95% confidence interval for each preprocessing combination applying the SVM model. 79
Table 20 – The results achieved by the proposed model extracting a single set of topics from the 197 documents. The best accuracy achieved for each scenario is highlighted in bold. 80

Table 21 – The RMSE computed for each regressor considering a single topic set. . . 81
Table 22 – The results achieved by the proposed model extracting a distinct topic set for each organizational department. The best accuracy achieved for each scenario is highlighted in bold. 81
Table 23 – The accuracy of the NERC approach in identifying PII entities associated with each group. 87

LIST OF ACRONYMS

ABNT	Associação Brasileira de Normas Técnicas
ACM	Association for Computing Machinery
BERT	Bidirectional Encoder Representations from Transformers
BoW	Bag-of-Words
CCPA	California Consumer Privacy Act
CNN	Convolutional Neural Network
CRF	Conditional Random Fields
DL	Deep Learning
DLP	Data Leakage Prevention
DT	Decision Tree
EC	Exclusion Criterion
EO	Executive Order
GB	Gradient Boosting
GDPR	General Data Protection Regulation
GPR	Gaussian Process Regression
HIPAA	Health Insurance Portability and Accountability Act
HTML	Hypertext Markup Language
IC	Inclusion Criterion
IEEE	Institute of Electrical and Electronics Engineers
ISMS	Information Security Management System
ISO	International Organization for Standardization
LASSO	Least Absolute Shrinkage and Selection Operator
LDA	Latent Dirichlet Allocation
LGPD	General Data Protection Law (Lei Geral de Proteção de Dados Pessoais, in Portuguese)
LSI	Latent Semantic Indexing
LSTM	Long Short-Term Memory
NER	Named Entity Recognition
NERC	Named Entity Recognition and Classification
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
NN	Neural Networks
OCR	Optical Character Recognition

OECD	Organization for Economic Cooperation and Development
PDET	Personal Data Entity Types
PDF	Portable Document Format
PHI	Protected Health Information
PII	Personally Identifiable Information
PoS	Part-of-Speech
RE	Relation Extraction
RF	Random Forest
RFC	Request for Comments
RMSE	Root Mean Square Error
RNN	Recursive Neural Network
RPC	Rate of Perplexity Change
RSLP	Removedor de Sufixos da Lingua Portuguesa (in Portuguese)
SLR	Systematic Literature Review
SPI	Sensitive Personal Information
SQL	Structured Query Language
SVM	Support Vector Machine
SVR	Support Vector Regressor
TF-IDF	Term Frequency-Inverse Document Frequency
US	United States
USD	United States Dollar
XML	Extensible Markup Language

CONTENTS

1 INTRODUCTION 13

2 BACKGROUND 17
- 2.1 Information Classification . . . 17
- 2.2 Natural Language Processing and Text Mining . . . 23
- 2.3 Information Value . . . 26

3 RELATED WORK 30
- 3.1 Selection Process . . . 30
- 3.2 Performing the Research Protocol . . . 32
- 3.3 Discussion of Information Classification Studies . . . 33
- 3.4 Discussion of NERC Studies . . . 37
- 3.5 Related Work Analysis and Research Opportunities . . . 39

4 PROPOSED MODEL 42
- 4.1 Personal and Context Values . . . 42
- 4.2 Entity Analysis . . . 45
- 4.3 Topic Analysis . . . 45
- 4.4 Information Value Computation . . . 46
- 4.5 Information Classification . . . 47

5 MATERIALS AND METHODS 48
- 5.1 Experiments Corpora . . . 49
  - 5.1.1 Corpora and PII Entities for Experiment 1 . . . 50
  - 5.1.2 Corpus for Experiments 2 and 3 . . . 52
- 5.2 Experiment 1: PII Discovery Evaluation . . . 53
  - 5.2.1 Annotated Entities . . . 53
  - 5.2.2 Preprocessing . . . 54
  - 5.2.3 Selection of NERC Approaches . . . 55
  - 5.2.4 Post-processing and Relation Extraction . . . 56
  - 5.2.5 Experiment Evaluation . . . 57
- 5.3 Experiment 2: Document Classification . . . 58
- 5.4 Experiment 3: Information Classification . . . 60
  - 5.4.1 Document value estimation . . . 60
  - 5.4.2 Features extraction . . . 64
  - 5.4.3 Training the regression model . . . 65
  - 5.4.4 Model Evaluation . . . 66

6 RESULTS AND DISCUSSIONS 69
- 6.1 Experiment 1: PII Entities Discovery . . . 69
  - 6.1.1 NERC Models Performance . . . 69
  - 6.1.2 Relation Extraction for PII Discovery . . . 71
  - 6.1.3 Undiscovered Documents . . . 73
  - 6.1.4 Discussion and Limitations . . . 75
- 6.2 Experiment 2: Document Classification . . . 77
- 6.3 Experiment 3: Information Classification Model . . . 79

6.3.1 The Proposed Model using a Single Topic Set 79
6.3.2 The Proposed Model using a Topic Set by Department 80
6.3.3 Discussion and Limitation 87

7 CONCLUSION 91

REFERENCES 94

1 INTRODUCTION

The digital transformation has changed how companies produce and sell their products and services. The basis of digital transformation is information, and organizations must protect them against cyber threats (MÖLLER, 2020). Information is considered an intangible asset and holds an economic value for organizations (FOSTER; CLOUGH, 2018). However, an actual challenge for organizations is discovering where the information is stored (BUNKER, 2012) and how to protect it (BERGSTRÖM; LUNDGREN; ERICSON, 2019). An additional challenge emerges from the uneven data value to organizations, so understanding what information is relevant to an organization is necessary to protect it cost-effectively (BUNKER, 2012). Ultimately, BERGSTRÖM; LUNDGREN; ERICSON (BERGSTRÖM; LUNDGREN; ERICSON, 2019) mentions that determining information value is a complex and problematic task.

An approach to distinguish information relevance according to its sensitivity is applying the information classification, described as the process of defining the sensitivity level of information to ensure that the proper protection is assigned to it, considering the information value, legal requirements, sensitivity, and criticality (ABNT, 2020). The information classification process is used by an organization to assign a persistent label to information assets in order to ensure that they are appropriately managed (NEWHOUSE et al., 2023). Information classification plays a central role in information security since it offers visibility about the sensitivity level of information, allowing security personnel to apply appropriate security controls. In this study, we also refer to information classification as security classification since academic literature applies this term frequently.

Despite the relevance of information classification, deploying it in practice has many challenges. One reason is that a person (named information owner) must be responsible for the information classification (ABNT, 2020), and a human-based approach introduces complexity to everyday practices. For example, the lack of skills of the information owners negatively impacts information classification implementation because they can act differently from the expected, resulting in the incorrect classification (NIEMIMAA; NIEMIMAA, 2017).

An additional challenge is the subjectivity involving the process, and the consequence is that different information owners classify the same or similar information in distinct sensitivity levels, also leading to inconsistent classification (BERGSTRÖM; ANTERYD; ÅHLFELDT, 2018; AKSENTIJEVIĆ; TIJAN; AGATIĆ, 2011; ANDERSSON, 2023). The information classification scheme is also a concern because each organization can define a specific number of levels (SINGH; RISHIWAL; KUMAR, 2018), and models need to adapt to distinct schemes. A final challenge is the necessity to reclassify information always that it changes (BUNKER, 2012). In a human-based approach, the information owner must check if the sensitivity level did not change after a modification. For example, an innocuous document should have its classification level elevated if someone inserts confidential information.

The barriers related to the human-based approach have gained research attention, and many studies evaluated distinct proposals for automating the information classification process (HEINTZ et al., 2022; LIANG et al., 2019). Automation demands a technique to recognize document features to identify their sensitivity level. Data Leakage Prevention (DLP) studies also addressed information classification since it can be used as a basis to authorize or deny an action related to a document. Natural Language Processing (NLP) can support information classification automation because documents are written in natural language. NLP is defined as the collection of computational techniques to analyze and represent human languages (CHOWDHARY; CHOWDHARY, 2020)

Studies have evaluated the application of NLP to support information classification. NEERBEK; ASSENT; DOLOG (NEERBEK; ASSENT; DOLOG, 2018) proposed an approach to identify sensitive information in electronic documents. The study argued that state-of-the-art approaches ignored the context to assume whether the information is sensitive. The proposal used a Recursive Neural Network (RNN) to learn about phrase structures associated with sensitive information to predict documents' sensitivity. The study of LIANG et al. (LIANG et al., 2019) analyzed the application of incremental learning to classify documents considering the changes affecting information over time. In a distinct approach, GEETHA; KARTHIKA; KUMARAGURU (GEETHA; KARTHIKA; KUMARAGURU, 2021) proposed a model to classify information according to the presence of Personally Identifiable Information (PII).

Unlike GEETHA; KARTHIKA; KUMARAGURU (GEETHA; KARTHIKA; KUMARAGURU, 2021), most information classification studies have not considered the presence of PII in the classification decision. The debate on privacy has increased significantly in recent years, and new laws and regulations to impose obligations on organizations about personal data have been issued. In Brazil, the federal government sanctioned a specific bill comprising data protection known, in Portuguese, as 'Lei Geral de Proteção de dados' (LGPD) (BRASIL, 2018). NLP can also support organizations in automating the discovery of PII in natural language-written documents. The NLP task recommended to identify PII is information extraction; more specifically, the sub-task referred to as Named Entity Recognition and Classification (NERC).

Scholars have evaluated the application of NERC for discovering PII entities. The study of DIAS et al. (DIAS et al., 2020) evaluated using shallow classifiers to identify PII in European Portuguese documents. In recent years, research has focused on deep learning-based NERC implementations (LI et al., 2020). YAYIK et al. (YAYIK et al., 2022) applied a Bi-LSTM classifier to recognize personal data in Turkish texts, and PETROLINI; CAGNONI; MORDONINI (PETROLINI; CAGNONI; MORDONINI, 2022) assessed two BERT-based models to identify sensitive data.

The aforementioned studies offered alternatives to automate information classification, yet most defined a strict number of sensitivity levels and performed the process using supervised learning. Supervised approaches usually offer significant results in a specific context (i.e., an organization). However, they are primarily dependent on corpora comprised of a massive number of documents for training (OTHMAN; FAIZ; SMAÏLI, 2022), and human labeling is time-consuming (COLACE et al., 2014). Furthermore, supervised learning poses challenges to the information classification process. For example, a new model must be retrained if an information classification scheme is updated or due to the discovery or creation of new valuable documents.

Regarding the NERC approaches mentioned, studies in the area analyzed the use of machine learning-based NERC to identify coarse-grained and well-known entities (PEARSON; SELIYA; DAVE, 2021; KAPLAN, 2020; DIAS et al., 2020), and we believe it is essential to evaluate machine learning approaches in fine-grained discovering. Additionally, only a few studies considered Relation Extraction (RE), another information extraction sub-task, and it is essential to minimize false-positives occurrences. Finally, both information classification and NERC areas lack studies evaluating NLP application in Brazilian Portuguese (SOUZA et al., 2018), which indicates the importance of new research focusing on the language. We did not find other studies on information classification and PII discovery targeting Brazilian Portuguese documents and Brazilian PII entities during this research.

We propose a distinct model to overcome the supervised learning barriers, which also considers personal data in the information classification process. The proposed model estimates the information value and assigns the sensitivity level based on it. We considered two aspects for estimating the value: the presence of PII in a document and the content context. However, according to BENDECHACHE et al. (BENDECHACHE et al., 2023), the information value comprises a wide range of definitions across different domains. So, the value-based approach proposed in this study adopts the Market-Based Model that is part of the data valuation framework proposed by FLECKENSTEIN; OBAIDI; TRYFONA(FLECKENSTEIN; OBAIDI; TRYFONA, 2023). The Market-Based Model includes assessing the information value based on data breach or loss costs, and the proposed model uses these costs to estimate a document's worth in monetary terms.

Studies have applied the information value approach to address challenges associated with different research topics in the information security area, such as compliance (DOHERTY; TAJUDDIN, 2018) and risk assessment (IBNUGRAHA; NUGROHO; SANTOSA, 2020). However, to the best of our knowledge, scholars have not evaluated the value-based approach for automating information classification. This work addresses this information classification research gap by answering the question: "Do applying NLP tasks allow extracting PII and context features to estimate information value and classify information?". This thesis hypothesis is that it is possible to associate internal and external costs to a sample of documents and employ different NLP tasks to extract document features, which allows the assessment of the value of new documents.

Aiming to respond to the research question, this study proposes an NLP-based model that evaluates the application of information extraction and topic modeling for obtaining document features and applies a regression model to estimate the document value. We understand that these NLP tasks can extract features related to the two aspects considered for estimating a document value. Information extraction discovers PII entities, indicating the presence of PII in a document, and topic modeling retrieves the more relevant topics in the document to represent its context. We based the model assessment on the experimental method by designing three experiments. The first experiment evaluates the performance of the information extraction subtasks to identify Brazilian-related PII. A second experiment analyzes the performance of a Bag-of-Words (BoW) approach to classify documents into different organizational departments. The last experiment evaluates the model's performance to estimate the document's value considering PII and context features. We crafted distinct corpora for evaluating the experiments based on documents provided by organizations in the educational and health industry sectors. The present study contributes as follows:

- Pioneers an information classification model based on the information value to determine the sensitivity level of a document. This approach allows organizations to change the number of sensitivity levels in an information classification scheme without retraining a machine learning model.

- Proposes a model that considers the presence of PII entities and information context to security classify a document.

- Pioneers the evaluation of NERC models to identify fine-grained entities related to Brazilian PII.

- Experiments an NLP-based implementation for automating information classification using Brazilian Portuguese documents.

The remainder of this work is organized as follows. Chapter 2 introduces the primary concepts to allow the proposal understanding. Chapter 3 discusses state-of-the-art studies associated with information classification and PII discovery. Chapter 4 presents the proposed model and its contributions. Chapter 5 describes the experiments to validate the proposed model. Chapter 6 presents and discusses the results achieved in the experiments. Finally, Chapter 7 summarizes the work contribution.

2 BACKGROUND

This work proposes the application of NLP to automate the information classification process. As the proposal integrates different research fields, this chapter presents relevant concepts and terminology from both areas to improve the understanding of the work. In order to cover the topics, we divided this chapter into three sections. Section 2.1 introduces information classification and presents the academic research related to the area. Section 2.2 details the NLP process and introduces the tasks associated with the proposed model. Finally, Section 2.3 describes academic proposals to estimate information value.

2.1 Information Classification

Information security is defined as the protection of information and information systems to avoid unauthorized access and use, preserving information from disclosure, modification, and disruption (GUTTMAN; ROBACK, 1995). SOLMS; NIEKERK (SOLMS; NIEKERK, 2013) described that information is the asset protected by information security, regardless of whether the asset is digital or physical. The scope of information security is different from cybersecurity, defined by ABNT (ABNT, 2015) as security in cyberspace. ABNT (ABNT, 2015) explains cyberspace as a complex environment associated with people, software, and services on the internet, connected through devices and networks that do not exist physically.

SOLMS; NIEKERK (SOLMS; NIEKERK, 2013) establishes an overlap between information security and cybersecurity, declaring that both must address information stored and transmitted electronically. As this thesis proposes the application of NLP to automate information classification and we consider only electronic information, its scope is related to this overlapped area comprising cybersecurity and information security. Therefore, we associate information classification with both research fields.

To attain information security, organizations must ensure confidentiality, integrity, and availability properties across various stages of the information lifecycle (ISO, 2018). According to ISO (ISO, 2018), confidentiality assures that only authorized entities can access information; integrity ensures the accuracy and completeness of information; availability makes the information available and usable by authorized entities. These properties are known as the security triad, and an organization needs to assure them to guarantee the information value (WHITMAN; MATTORD, 2017).

In order to protect the information, an organization must plan, implement, and manage appropriate security controls by establishing an Information Security Management System (ISMS) (ISO, 2018). Security control is an action, device, procedure, technique, or other measures to reduce the existing vulnerabilities of an information system, protecting confidentiality, integrity, and availability of the information (PAULSEN; BYERS, 2019). ABNT (ABNT, 2022) lists four security control sets comprising 93 security controls that an

ISMS should consider based on an organization's requirements.

As part of the organizational control set, information classification is the security control that establishes the relevance of information and applies a classification label to offer information sensitivity visibility. According to ABNT (ABNT, 2020), information classification defines the sensitivity level of information. NEWHOUSE et al. (NEWHOUSE et al., 2023) present information classification as a process organized into five activities. From the definition of a classification scheme, which establishes the classification levels, an organization identifies its information assets and determines at which level the information must be classified (NEWHOUSE et al., 2023). The following steps are to assign a label to information and monitor if its classification must be changed (NEWHOUSE et al., 2023). The classification label aims to communicate the sensitivity level of information (ABNT, 2022). The classification label offers visibility, allowing an organization to analyze the information sensitivity and the risks associated with each type of information asset (STATES, 2019).

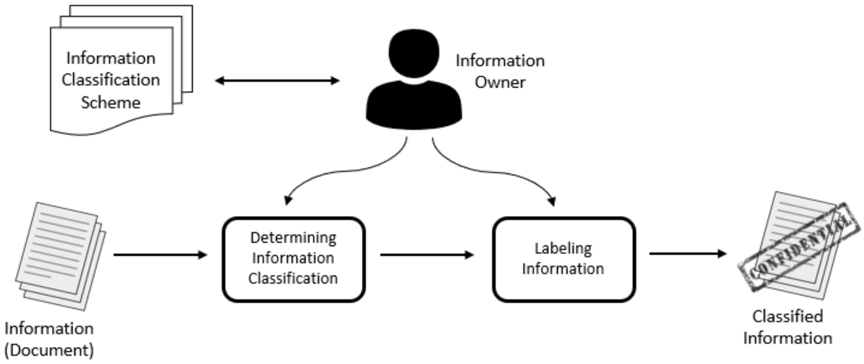
Information classification is necessary because not all information must be protected (PELTIER, 1998). Instead, the application of security controls depends on identifying sensitive information, and this is a challenging activity for many organizations (BUNKER, 2012). New business trends and technologies have brought additional difficulties in identifying and protecting sensitive information (MORROW, 2012; SHA et al., 2018). Digital transformation and new technologies have increased the volume of data generated, and a significant part of the information produced is valuable to the organization that generated it (REINSEL; GANTZ; RYDNING, 2018).

According to REINSEL; GANTZ; RYDNING (REINSEL; GANTZ; RYDNING, 2017), approximately 87% of new information produced in 2025 will demand some security control. The study surveyed information technology and security professionals from 17 countries and reported that the biggest challenge in the cryptographic security control application is discovering where sensitive data resides in the organization (PONEMON, 2022). This type of information demonstrates that security classification is an actual challenge today, and the rise of data can make this challenge even more significant.

Although information classification is a well-known security control, its implementation demands organization resources to identify the value of different types of information. An information classification scheme must specify the classification level formed by categories representing information sensitivity (ABNT, 2020). The standard also defines the role of the information owner, who analyzes and assigns each electronic document's classification level. So, information owners should analyze information and identify its value to apply the correct classification.

Figure 1 depicts an overview of a human-based information classification process. In the figure, a new document has to be classified, so the information owner needs to analyze the information and verify the proper sensitivity level in the information classification scheme. After determining the sensitivity level, the information owner must assign the label that makes

Figure 1 – An overview of the information classification process.



Source: Created by the author.

the information classification visible. The information classification scheme is part of the information classification policy, which presents the guidelines regarding information classification in an organization (NEWHOUSE et al., 2023). Table 1 presents an example of an information classification scheme from the Whirlpool Corporation, which defines three sensitivity levels (NICOLLS, 2002).

Table 1 – An example of the three-level information classification scheme from the Whirlpool Corporation.

Sensitivity Level	Description
Confidential	A subset of Whirlpool Internal information, the unauthorized disclosure or compromise of which would likely have an adverse impact on the company's competitive position, tarnish its reputation, or embarrass an individual. Examples: Customer, financial, pricing, or personnel data; merger/acquisition, product, or marketing plans; new product designs, proprietary processes and systems.
Internal	All forms of proprietary information originated or owned by Whirlpool, or entrusted to it by others. Examples: Organization charts, policies, procedures, phone directories, some types of training materials.
Public	Information officially released by Whirlpool for widespread public disclosure. Example: Press releases, public marketing materials, employment advertising, annual reports, product brochures, the public web site, etc.

Source: (NICOLLS, 2002)

It is essential to highlight that information classification is not restricted to organizations. Countries also have concerns about the types of information that need to be protected. In 2009, the United States (US) issued the last Executive Order (EO) concerning information classification (PRESIDENT, 2009). EO 13526 claims that keeping some information secret is required to assure national defense and protect citizens, democratic institutions, homeland security, and interactions with foreign nations. The EO 13526 also determines the use of classification to protect information adequately.

The United Kingdom also addressed information classification, establishing a policy to describe how the government must classify information assets (OFFICE, 2018). The policy applies to all information collected and handled by the government, including information associated with foreign partners. The document highlights that information "has intrinsic value and requires an appropriate degree of protection" (OFFICE, 2018). Brazil also defined guidelines concerning information classification in a national law (BRASIL, 2011). The law presents reasons for the government to classify information. National defense, international agreements, and citizen's lives are some of them. The law also establishes the criteria to declassify Brazilian national information.

The academic literature uses different terms for information classification, such as data classification and security classification. Although some authors point out slight differences about some terms (BERGSTRÖM, 2017), we consider that all terms aim to identify the relevance of the information to apply proper security controls. In the 1970s, the seminal study of BELL; LAPADULA (BELL; LAPADULA, 1996) established an access control system based on classification. The proposed system classifies subjects and objects, assigning a clearance level to subjects and a classification level to objects. When a subject requests access to an object, the system verifies both security levels and authorizes or denies the action. In another fundamental study regarding information classification, DENNING (DENNING, 1976) discussed security classes associated with the information in a study about access control mechanisms to restrict information exchange on a network.

1980s studies focused mainly on using information classification by governments (GARFINKEL, 1984; HALPERIN, 1984). In contrast, the 1990s works addressed its application to organizations. ELOFF; HOLBEIN; TEUFEL (ELOFF; HOLBEIN; TEUFEL, 1996) proposed an approach to classify electronic documents considering security criteria regarding the documents' business lifecycle. In another work, PELTIER (PELTIER, 1998) presented guidelines and good practices to implement an information classification policy on business. Since the 1990s, studies have discussed the need to declassify and reclassify information automatically (PELTIER, 1998). At that moment, organizations were intensifying the use of technology and connecting their infrastructure to the internet, so organizations realized the importance of information classification to protect their information. Therefore, the publications focused on aiding the companies in implementing information classification correctly.

In recent years, information classification has been considered a well-known security control and has been relatively adopted by distinct organizations and government sectors. This landscape stimulated new approaches to improve how information classification is implemented in the real world and intensified the integration with additional business processes (SIPONEN; BASKERVILLE; KUIVALAINEN, 2005). An example is the study of FARN; LIN; LO (FARN; LIN; LO, 2008) that, based on recognized standards, presented a new framework to classify information on the system called e-Taiwan. Novel technologies also provided additional challenges to the information security domain, and studies discussed the application of information classification in cloud computing platforms (ONWUBIKO, 2010) and the internet of things (LU et al., 2014).

In the later 2000s', studies started to discuss the automation of information classification (HENNESSY et al., 2009; MATHKOUR; TOUIR; AL-SANIE, 2005). Researchers leveraged artificial intelligence resurgence and assessed its integration with text mining to automatically assign security labels to unstructured content (BROWN; CHARLEBOIS, 2010). At this point, scholars started evaluating several approaches for using text mining to classify information based on its sensitivity. ALNEYADI; SITHIRASENAN; MUTHUKKUMARASAMY (ALNEYADI; SITHIRASENAN; MUTHUKKUMARASAMY, 2013) proposed the creation of profiles based on N-gram analysis and used word frequencies to determine the relevance of documents. First, each profile was associated with a classification level, and next, each document was assigned a profile.

In a distinct idea, THORLEUCHTER; POEL (THORLEUCHTER; POEL, 2012) defined a highly granular approach and used Latent Semantic Indexing (LSI) to assign classification levels to distinct parts of a document. The study split a document into several objects and assigned a classification level for each object. This approach allowed the users access to document objects according to their permissions. Thus, even a user who does not have permission to access parts of the document can read the authorized ones. ZARDARI; JUNG (ZARDARI; JUNG, 2016) also associated an electronic document with different classification levels. The study considered two security classes (confidential and non-confidential), and it proposed a new version of the K-Nearest Neighbors classifier to predict which information in a document corresponds to each class.

The increasing use of artificial intelligence and machine learning has aroused discussions about using "black boxes" classifiers, which do not offer a human-understandable explanation about how the result was achieved (GOEBEL et al., 2018). A study evaluated using a machine learning classifier that offers a human interpretable result to perform information classification (ENGELSTAD et al., 2015a). The study assessed the Least Absolute Shrinkage and Selection Operator (LASSO) classifier based on two or multiple classification levels. The authors highlighted that the classifier could generate a list of human-understandable words used to predict the classification level, which offers an accuracy that makes information classification feasible.

Another concern addressed in academic papers was using real data to analyze the applicability of automated information classification in the real world. As ALZHRANI et al. (ALZHRANI et al., 2016b) mentioned, many approaches were evaluated using artificially generated sensitive documents, making the analysis less accurate. The study of HUANG; CHIANG; CHANG (HUANG; CHIANG; CHANG, 2018) proposed a system to assign the classification level to corporate email, aiming to avoid confidential data leakage in electronic messages. The system deals with imbalanced classes by applying the clustering task to perform undersampling. The experiment used real-world email data provided by a world-leading company in Taiwan. ALZHRANI et al. (ALZHRANI et al., 2016b) also used a real-world dataset to evaluate an automated information classification approach. The authors proposed using Latent Dirichlet Allocation (LDA) to improve the training set used by a machine learning classifier. The authors proposed this improvement to adjust the classification process to a big data scenario. The experiment used sensitive US diplomatic cables leaked on Wikileaks in the evaluation.

Academic literature also studied the implementation of information classification in the real world and pointed out that organizations have faced many challenges. BERGSTRÖM; ÅHLFELDT (BERGSTRÖM; ÅHLFELDT, 2014) enumerated seven challenges associated with applying information classification in different organizations. One challenge is that the human factor can contribute to inconsistent classification because information owners can have different perceptions about the sensitivity of the information included in a document (BERGSTRÖM; ÅHLFELDT, 2014). Moreover, an organization can assign the information owner's role to dozens or hundreds of employees. In this case, an additional challenge is establishing objective criteria to support information owners in deciding which classification level they must assign to a document. The lack of explicit process guidance and criteria to decide the value of information can result in under or overclassified information (BERGSTRÖM; KARLSSON; ÅHLFELDT, 2021)

Another challenge mentioned in BERGSTRÖM; ÅHLFELDT (BERGSTRÖM; ÅHLFELDT, 2014) was defined as performance management. This challenge involves the complexity of identifying which information in an electronic document is relevant to the organization for estimating the document's value. As already discussed, this situation generates inconsistent classification. ANDERSSON (ANDERSSON, 2023) also mentioned the inconsistency in classification due to the subjective judgment by the information owner in a study presenting five challenges for implementing the information classification process. KAARST-BROWN; THOMPSON (KAARST-BROWN; THOMPSON, 2015) interviewed 29 employees working in distinct departments from five different industries. The authors affirmed that employees had no problem distinguishing sensitive from non-sensitive information. However, the study's findings corroborated previous research and asserted that employees demonstrated difficulty comprehending the sensitivity level associated with information.

BERGSTRÖM; ÅHLFELDT (BERGSTRÖM; ÅHLFELDT, 2014) mentioned that another difficulty related to performance management is the need to keep the classification level updated. If an electronic document suffers some modification, the information owner should evaluate if the classification level remains unchanged or needs revision. In a study encompassing Swedish government agencies, BERGSTRÖM; ANTERYD; ÅHLFELDT (BERGSTRÖM; ANTERYD; ÅHLFELDT, 2018) concluded that organization struggles to reclassify information, and the lack of mechanisms for detecting changes in information over time made the reclassification a challenging activity.

The inconsistency and the reclassification challenges can negatively impact the adoption of information classification for organizations. However, an automated model uses an objective set of criteria to define the classification level of information. Additionally, the automation can deal with the reclassification challenge since it can monitor document modification and reevaluate the classification level. Information classification automation demands techniques to deal with unstructured documents, and the research areas addressing this format are NLP and text mining, introduced in the following section.

2.2 Natural Language Processing and Text Mining

Text mining applies multiple tasks to use information from documents to reach an objective based on insights extracted from the documents' content (AGGARWAL; ZHAI, 2012). Text mining is a process to extract valuable knowledge from data sources, similar to data mining. However, unlike data mining, the data sources used in text mining are formed by document collections, and the process extracts knowledge from unstructured data present in the documents (FELDMAN; SANGER et al., 2007). The unstructured data analyzed by text mining usually comprises texts written in a natural language (JO, 2018).

NLP applies linguistic concepts to extract a meaning representation from a text (KAO; POTEET, 2007). NLP comprehends some techniques that can be applied to understand natural language content and make inferences based on this capability (ZHAI; MASSUNG, 2016). Although some authors understand that NLP is the research area responsible for extracting knowledge from the text (VAJJALA et al., 2020), this work considers NLP as part of the text mining area (MINER et al., 2012). While NLP focuses on natural language-written texts, text mining encompasses a broader range of document formats, including semi-structured ones, like HTML and XML files (YANG; CHEN, 2002). This study focuses on classifying information in documents written in natural language, so it relates to both research fields.

NLP offers a collection of fundamental tasks that allow distinct text-based operations to be carried out in documents, resulting in different types of knowledge. VAJJALA et al. (VAJJALA et al., 2020) arranged NLP tasks according to their difficulty in implementing and classified the three tasks utilized in this proposal as medium. The three tasks are information extraction, text classification, and topic modeling. In addition to these, other NLP tasks are information retrieval, text clustering, text summarizing, and sentiment analysis (GUPTA; LEHAL et al., 2009; WEIGUO et al., 2005; ALLAHYARI et al., 2017).

In order to extract knowledge from text, MINER et al. (MINER et al., 2012) describes a core process divided into three stages : (i) Corpus definition, (ii) Preprocessing, and (iii) Knowledge extraction. The first stage, corpus definition, collects relevant documents about a specific issue. The second stage comprehends the application of one or more preprocessing techniques. Preprocessing is responsible for transforming raw text data into an intermediate format, improving the extraction of patterns by a task (MUNKOVÁ; MUNK; VOZÁR, 2013;

ANGIANI et al., 2016; UYSAL; GUNAL, 2014). Preprocessing is required to improve the quality and usability of unstructured data. Usually, a unique preprocessing technique is not enough to prepare the text for processing, and it is pretty common to implement different techniques sequentially (FELDMAN; SANGER et al., 2007). There are several techniques for preparing unstructured data, and the decision on which to select depends on the task or tasks employed in the process.

One common approach to represent a document for knowledge extraction is BoW, which maps the document into a fixed-length vector (ZHAO; MAO, 2017). The BoW approach analyzes the frequency of terms in a document, often a word. The preprocessing technique that breaks up the text document into words or tokens is called tokenization (MINER et al., 2012). An alternative to a single word frequency is N-gram, which tokenizes groups of "n" consecutive words (ANANDARAJAN; HILL; NOLAN, 2019). In cases where the BoW approach calculates each word's frequency, a preprocessing technique is necessary to remove frequent words that add no value to the analysis, such as prepositions and articles. These words are named stopwords, and the technique is named stop word removal (ANANDARAJAN; HILL; NOLAN, 2019).

One challenge in text mining analysis is dealing with word inflection. An alternative is a stemming function that reduces variants of words to the root form. A more complex option is lemmatization, which performs morphological analysis of each word to determine the lemma (SINGH; GUPTA, 2017). Preprocessing can also address language syntax by applying the Part-of-Speech (PoS) technique, which identifies the syntactic role for each word in a sentence (EISENSTEIN, 2018). The last stage of the process involves extracting the knowledge from preprocessed electronic documents. The tasks can vary according to the type of knowledge targeted. In this section, we focused on detailing the NLP tasks related to the proposed model.

A well-known task is text classification, a supervised learning approach that matches new observations according to a dataset properly categorized and labeled (MILOSEVIC; DEHGHANTANHA; CHOO, 2017). The dataset contains a set of documents, each associated with a class value. This dataset is referred to as training data, and it is necessary to build a classification model. The model associates document features with one or more classes and predicts the class of a new set of documents, named testing data (AGGARWAL; ZHAI, 2012). Text classification can use shallow or deep learning. Shallow learning is based on traditional classifiers like Naive Bayes, Decision Tree, and Random Forest and employs statistical models (KUMAR et al., 2020). Shallow learning demands manual feature extraction to train and evaluate the classifier (ALTINEL; GANIZ, 2018). Deep Learning (DL) explores robust Neural Networks (NN) that automatically learn high-level features from a text document, dispensing manual feature extraction (DENG; LIU, 2018).

Information extraction aims to discover structured information from unstructured content (AGGARWAL; ZHAI, 2012), and it is usually performed through two subtasks: Named Entity Recognition (NER) and RE. These subtasks identify known entities like people, companies, and places in a text and infer relationships among them (GUPTA; LEHAL et al., 2009). Initially, NER was developed as a rule-based system, but state-of-the-art implementations rely on statistical machine learning methods (AGGARWAL; ZHAI, 2012). A recent approach transformed NER into a classification problem and used supervised learning to identify the entities. The approach is called NERC (NADEAU; SEKINE, 2007).

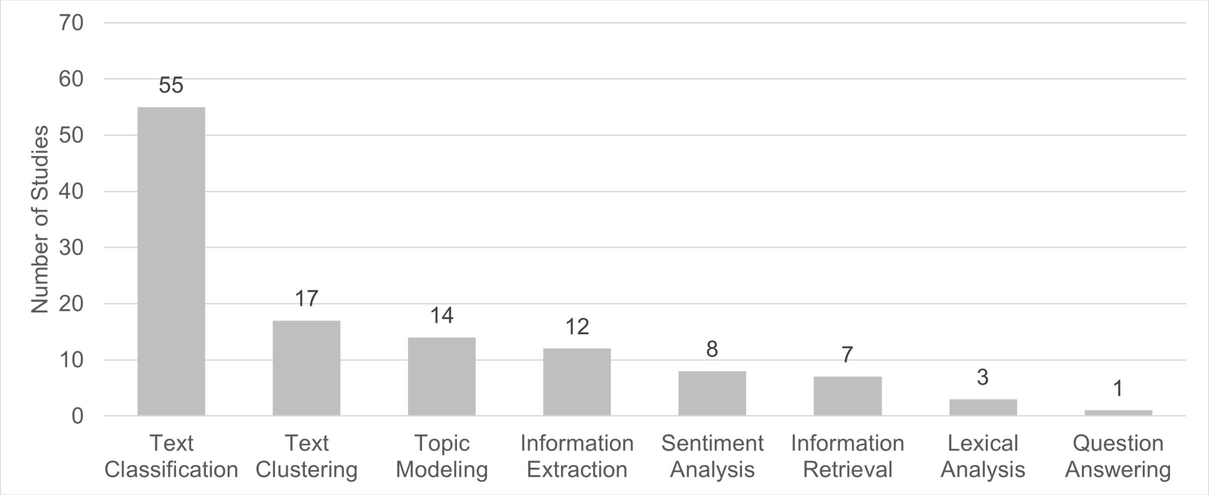
After implementing NER for entity identification, the next step is determining how they are related (EISENSTEIN, 2018). RE aims to detect and characterize semantic relations between identified entities (AGGARWAL; ZHAI, 2012). YAN; SUN; LIU (YAN; SUN; LIU, 2021) described a RE system based on three activities: (i) the identification of entities with specific meanings in the text, (ii) the recognition of the relations present in the text, and (iii) the development of connections linking the recognized entities and relations. The study of DETROJA; BHENSDADIA; BHATT (DETROJA; BHENSDADIA; BHATT, 2023) categorized RE into traditional and DL methods. The traditional ones were divided into rule-based and machine-learning methods, while DL approaches were broken into supervised learning and distant supervision methods.

Topic modeling is an unsupervised learning that uses statistical methods for analyzing texts and discovers the main themes pervading a collection of documents (BLEI, 2012). These themes are called topics and consist of word clusters representing the ideas discussed in text data. Topic modeling uses the co-occurrence of topics to identify the subject associated with documents (JACOBI; ATTEVELDT; WELBERS, 2016). The most frequently used topic modeling algorithm is LDA, which implements the BoW approach (VAYANSKY; KUMAR, 2020). It is important to highlight that LDA instability is a known concern and should be considered by researchers (AGRAWAL; FU; MENZIES, 2018). Current algorithms aim to improve the topic discovery in specific analyses, such as short text and time-based topic modeling (VAYANSKY; KUMAR, 2020).

The predominance of NLP tasks changes according to the research area. Text classification is the prevalent task in studies related to psychiatry (ABBE et al., 2016), education (FERREIRA-MELLO et al., 2019), and identifying papers for inclusion in systematic reviews (O'MARA-EVES et al., 2015). On the other hand, a study in innovation research presented text clustering as the most used task (ANTONS et al., 2020), and information extraction is the most common in the agricultural domain (DRURY; ROCHE, 2019).

We conducted a Systematic Literature Review (SLR) comprising the application of text mining in the cybersecurity domain that analyzed 83 studies (IGNACZAK et al., 2021). The SLR findings corroborated to SOUZA et al. (SOUZA et al., 2018) and pointed out the lack of studies evaluating NLP tasks in Portuguese. Unlike occurs for the English Language, Portuguese is considered a low-resource language (NETO; SILVA; SOARES, 2023). In the SLR, we identified that 70 studies focused on the English language. Further, the studies only targeted five other languages: Arabic, Chinese, German, Russian, and Turkish. This result indicates the importance of new studies integrating NLP and cybersecurity focusing on

Figure 2 – Text mining tasks applied in the cybersecurity domain.



Source: Created by the author

Portuguese. Moreover, based on the SLR search string, we identified only three studies directly related to information classification (THORLEUCHTER; POEL, 2012; ALNEYADI; SITHIRASENAN; MUTHUKKUMARASAMY, 2013; ZARDARI; JUNG, 2016), highlighting a gap for new scientific studies in the area.

Considering the NLP tasks, the SLR identified text classification as the predominant task applied in the area, present in 55 studies. SLR also revealed that many studies used multiple tasks in the assessment. For example, considering the 55 studies that applied text classification to extract knowledge, 35 studies led an experiment that evaluated the task performance singly, whereas 20 studies used text classification combined with other text mining tasks. Figure 2 summarizes the number of studies associated with each task in the cybersecurity domain.

2.3 Information Value

An organization should protect the assets that hold some value for it. ROSS et al. (ROSS et al., 2018) characterize an asset as "an item of value to the organization's stakeholders". Assets can be divided into tangible and intangible (GRECO; CRICELLI; GRIMALDI, 2013). The Cambridge Dictionary defines a tangible asset as "a physical asset whose value can be easily measured" (DICTIONARY, 2021b). At the same time, it describes an intangible asset as "something valuable that a company has that is not material" (DICTIONARY, 2021a). As mentioned in 2.1, an ISMS addresses the implementation of security controls to defend information assets. Information assets are defined as "physical, hardware, software, data, communication, administrative and personnel resources of a computing system that once compromised will release sensitive system information to the threat agent" (VIDALIS, 2010).

The recent data explosion originated from big data changed how to assess the value of information (FOSTER; CLOUGH, 2018). Information, an intangible asset, has increased its value in the current global economy (OSINSKI et al., 2017) because digital transformation has boosted its use by organizations, making it even more critical. FOSTER; CLOUGH (FOSTER; CLOUGH, 2018) considered that information is a primary resource due to its financial worth, and it is also a secondary resource because it supports products and services provided by an organization.

The value of information is not a recent discussion. In the 60s, HOWARD (HOWARD, 1966) presented the theory for estimating information value to support decisions. The author discussed how much information is worth in a fictitious scenario where organizations compete for a contract and indicated that information can have an economic value. Based on the premise that information has a value, a relevant aspect is the definition of methods to estimate the value of specific information. HARRISON; SULLIVAN (HARRISON; SULLIVAN, 2000) proposed using qualitative or quantitative methods to estimate the value of intangible assets. Thus, the authors defined that qualitative measures are typically based on a judgment, and the value is absent in precise or quantifiable measurements. On the other hand, quantitative measures are expressed in a particular amount, such as dollars or other currencies.

The need to measure information value using quantitative methods has been discussed for a long time. In the 90s, MOODY; WALSH (MOODY; WALSH, 1999) highlighted the relevance of valuing information since it was recognized as a key economic resource and one of the most critical assets for a business. The authors also established seven laws proper of information, distinguishing it from other types of assets. We list the seven laws (MOODY; WALSH, 1999):

1. Information is shareable;
2. The value of information increases with use;
3. Information is perishable;
4. The value of information increases with accuracy;
5. The value of information increases when combined with other information;
6. More is not necessarily better;
7. Information is not depletable.

In the last decades, several studies have discussed methods to value information in different business areas, mainly based on the relevance of some information to make decisions. KEISLER et al. (KEISLER et al., 2014) surveyed several aspects of published articles to understand the use of the value of information. They encountered studies estimating information value in Agriculture, Economics, Energy, and Medical domains. The study also highlighted that the proposed methods measured the information value using a financial value, cost-benefit analysis, and multi-attribute utility functions.

THORNTON (THORNTON, 2013) grouped models and techniques to measure intangible assets in market, income, and cost. The market indicates the value of intangible assets based on the transaction price used to buy or sell similar assets in recent transactions. The authors highlighted that market distortions can influence the value even if an intangible has a quoted price. The income approach uses the expected future value of an asset to estimate its value. Despite their hypothetical characteristic, this approach is the most commonly applied to estimate the value of intangible assets. The last approach uses the cost required to repurchase or reproduce an asset due to physical deterioration, use, and obsolescence. The authors underlined that the cost approach may ignore the asset's future economic benefits, which is one of the reasons for its less widespread acceptance.

SAJKO; RABUZIN; BAČA (SAJKO; RABUZIN; BAČA, 2006) proposed two forms of quantitative assessment. A direct form comprises a piece of information with a defined financial value, such as licenses, patents, and projects. The indirect form involves types of information that do not have correspondent financial value. In these cases, the study proposed an assessment based on four components: (i) the value for a business, (ii) the value for other businesses, (iii) the cost of the reconstruction of the asset, and (iv) the time dimension. The time dimension comprises information devaluation over time.

In a perspective based on the user side, RAO; NG (RAO; NG, 2015) discussed the value of information in the big data scenario, where organizations collect users' information used to make money. The authors aim to clarify the value of information and incentivize users to demand monetary compensation for sharing their information. The definition of information value considers the understanding of the demand for some type of information by the market, considering that a part provides the information (user) and another is interested in buying it (buyer). According to the authors, the demand can be determined by three factors: (i) the amount of money that a buyer considers paying for the information, (ii) the number of interested buyers, and (iii) the velocity related to changes in the demand for the information.

In another study considering the selling of information by users, RAO; NG (RAO; NG, 2016) defined a price model using Shannon's information theory concepts. The authors considered the traditional economic concept that defines a higher price for a more demanded product discussed in their former study (RAO; NG, 2015). The price model calculates the demand price by interviewing information buyers. The study also applied information gain for every type of information based on previously defined structured data and determined the probability distribution associated with each type. The result is considered the information value of a specific information type multiplied by the demand price to estimate the information price.

The study of BENDECHACHE; LIMAYE; BRENNAN (BENDECHACHE; LIMAYE; BRENNAN, 2020) developed an approach to assess structured data value stored in a database automatically. The proposal also included a scoring method to rank the most valuable database tables. The authors initially proposed a survey addressing data value dimensions and applied it to business domain experts. Survey answers were used as a baseline for comparing results with the approach proposed. The approach initially performed data cleaning to remove inconsistencies and created a repository to store data assessed and results. The next step was to perform data valuation, in which the author designed an SQL query in line with each survey question. The authors compared the SQL query results to the survey answer, estimating the value of the databases based on the scoring method.

Despite the long-time discussion, in a recent survey, BENDECHACHE et al. (BENDECHACHE et al., 2023) affirmed that "there is currently no consensus on the definition of data value" and added that the definition changes according to the domain. For example, in machine-learning domains, data value refers to the relevance of the training sample or feature on impacting the model performance. In contrast, the data value estimation in the business domain is based on costs. The authors also mentioned that information value is highly dependent on the context. Through their examination, they identified eight relevant studies in the information security area; four focused on risk assessment, while the remaining four centered on privacy. The survey reinforces the originality of using information value definition to support information classification.

3 RELATED WORK

This thesis proposes a novel information classification model using NLP tasks to automate the process. Thus, we applied the SLR method proposed by KITCHENHAM; BUDGEN; BRERETON (KITCHENHAM; BUDGEN; BRERETON, 2015), which we applied in a prior SLR that analyzed the application of NLP tasks in the cybersecurity domain (IGNACZAK et al., 2021), to identify studies applying distinct NLP approaches to classify information according to its sensitivity. The SLR confirmed the growing interest in applying NLP in cybersecurity, besides the existing research gap related to automating information classification. Therefore, this chapter narrowed the focus and used the SLR method to present academic papers related to this research gap.

We also performed supplementary research to present studies discussing NERC implementations to recognize sensitive data in documents. As the proposed model uses NERC to identify personal data, presenting the research topic's state of the art is relevant to demonstrating contributions regarding using PII in the information classification process.

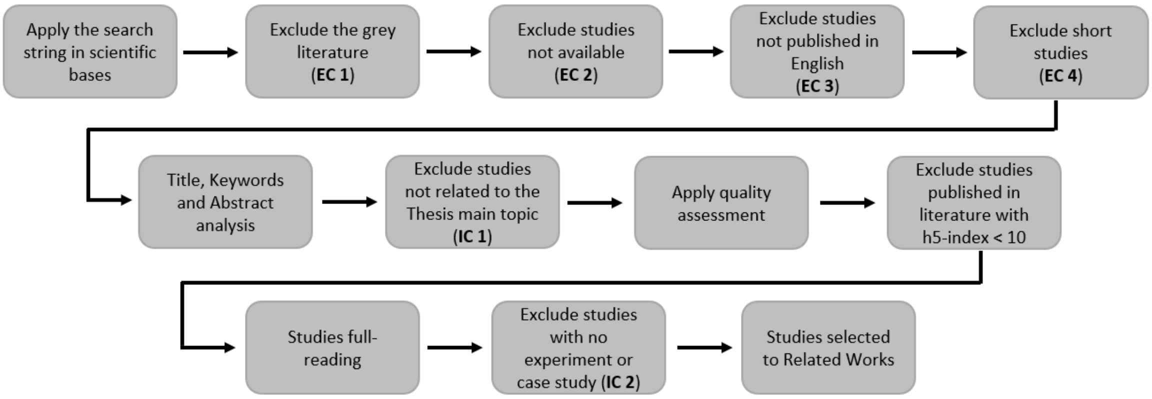
Aiming to offer an organized approach to the audience, we divided this chapter into five sections. Section 3.1 introduces the research protocol for selecting studies using NLP tasks to automate information classification. Section 3.2 describes the implementation of the research protocol, detailing the number of studies associated with each stage of the protocol. It is relevant to highlight that Sections 3.1 and 3.2 are unrelated to NERC implementations to discover personal data. In Section 3.3, we discuss the information classification studies selected using the SLR method. Section 3.4 shows the studies associated with NERC implementations. Finally, in Section 3.5, we analyze related work and highlight the research opportunities addressed in this study.

3.1 Selection Process

Papers selection followed the initial phase of the methodology for developing an SLR (KITCHENHAM; BUDGEN; BRERETON, 2015). The methodology's first phase, Plan Review, comprises tasks related to the research protocol design, such as the search strategy and the criteria for selecting the studies. Figure 3 shows the complete research protocol presenting the step-by-step process. We leveraged the experience of (BERGSTRÖM; ÅHLFELDT, 2014) to define the keywords used in the search string. The study analyzed the terms "information classification", "security classification", and "data classification" to design the search string. As the authors underlined that "data classification" generated many false positives, we decided not to use it.

In our study, we consider it relevant to analyze studies that evaluated the automation of information classification, so we added terms associated with NLP. Finally, we added the term "security" to get results associated with the security domain, such as "information security"

Figure 3 – The research protocol to select studies related to information classification.


Apply the search string in scientific bases
Exclude the grey literature (EC 1)
Exclude studies not available (EC 2)
Exclude studies not published in English (EC 3)
Exclude short studies (EC 4)
Title, Keywords and Abstract analysis
Exclude studies not related to the Thesis main topic (IC 1)
Apply quality assessment
Exclude studies published in literature with h5-index < 10
Studies full-reading
Exclude studies with no experiment or case study (IC 2)
Studies selected to Related Works

Source: Created by the author.

and "cyber security". Based on the study of BERGSTRÖM; ÅHLFELDT (BERGSTRÖM; ÅHLFELDT, 2014) and preliminary searches, we decided to use the context and keywords presented in Table 2. The search query used the logical operator "OR" between the keywords and the logical operator "AND" between the contexts.

Table 2 – Keywords used to discover relevant studies.

Context	Keywords
Information classification	information classification, security classification
Information security	security
Natural Language Processing	text mining, natural language processing, text classification

Source: Created by the author.

The search strategy limited the search query to some widely used scientific databases in computer science. The authors analyzed the sources used in the studies of (TRAN; ZDUN et al., 2017; MOUSTAKA; VAKALI; ANTHOPOULOS, 2018), and selected the following databases: ACM[1], IEEE Xplore[2], ScienceDirect[3], and Springer[4]. We performed the protocol search query in the four databases. The search comprised studies published from 2010 to the present.

[1] https://dl.acm.org/

[2] https://ieeexplore.ieee.org/

[3] https://www.sciencedirect.com/

[4] https://link.springer.com/

Table 3 – Criteria applied to include or exclude studies in the SLR.

ID	Inclusion/Exclusion Criteria
IC1	Study is related to the automation of information security classification or discovery of confidential information through an NLP task
IC2	Study must have an experiment or case study and present consistent results
EC1	Study not published in a journal or conference
EC2	Authors can not access the full paper
EC3	Study is not in English
EC4	Study is shorter than six pages

Source: Created by the author.

The research protocol specified by the authors also established criteria to exclude or include related work. We defined four exclusion criteria (EC) and two inclusion criteria (IC) shown in Table 3. After performing the search queries, the first task was to assemble all results in a single file and delete duplicated entries. These studies were considered candidate papers, and the next step is the application of exclusion criteria.

The first exclusion criterion aims to remove all grey literature because the authors decided to use only peer-reviewed studies. The second criterion deletes all candidate papers whose authors can not access directly from the databases chosen in the search strategy, and the third keeps just studies written in English. The last exclusion criterion removes short papers. The protocol defined two inclusion criteria to consider studies for discussion. The first criterion is an analysis based on the title and abstract to remove a paper if it is unrelated to the NLP application in information classification. The second criterion comprises a complete reading and analysis of whether the study includes a consistent experiment or case study related to information classification.

Another decision taken during the planning phase concerned the quality of the studies. KITCHENHAM; BUDGEN; BRERETON (KITCHENHAM; BUDGEN; BRERETON, 2015) defined the quality assessment as an essential part of the process because it can improve the value of SLR, allowing the reviewers to exclude papers that do not reach the expected quality level. Therefore, we decided to apply the quality assessment before the full reading. This study used the h5-index[5] score as a quality criterion and removed all papers published in proceedings or journals with a score minor than 10.

3.2 Performing the Research Protocol

After defining the research protocol, we implemented it step by step. We performed the search queries in the databases on February 1st, 2024, and found 117 candidate papers. Since

[5] https://scholar.google.com/intl/en/scholar/metrics.html#metrics

Table 4 – Summary of the criteria application to select the related works.

Databases	Studies returned	E1	E2	E3	E4	Excluded studies I1 - Round 1	Quality	I1 - Round 2	Selected Studies
ACM	12	0	0	0	0	11	1	0	0
IEEE Xplore	21	0	0	0	7	5	2	1	6
ScienceDirect	47	0	3	0	1	41	0	0	2
Springer	37	0	0	0	0	30	0	4	3
Total	**117**	**0**	**3**	**0**	**8**	**87**	**3**	**5**	**11**
Additional studies included									1
Total selected studies									**12**

Source: Created by the author

the searches did not return duplicated papers, we applied all studies to the exclusion criteria. The first and third criteria did not remove studies since all were published in journals or conferences and written in English. We could not access three studies, so we removed them based on the second criterion. Eight studies had less than six pages and were removed due to the fourth criterion.

The exclusion criteria removed 11 candidate papers, and the next step was to analyze the 106 remaining studies to verify if they fit the inclusion criteria. The authors analyzed the studies' abstract and title in the first round, and 87 candidate studies were removed. It was evident from the reading that they were unrelated to the automation of information classification.

The 19 selected studies in the first round were applied to the quality assessment. The protocol established the use of the h5-index as a quality threshold, and there is no reason for a detailed reading of a set of papers that the quality criterion will remove. All journals and conferences were applied to check the h5-index metric in February 2024. The quality assessment excluded three studies, so 16 candidate papers reached the next round. After the complete reading to perform the second inclusion criterion, the authors excluded five studies.

Although we analyzed several search strings to define the research protocol, a relevant study was not included. Since the study performed research similar to our proposal, we added it to related work. It is imperative to mention that the study was verified against exclusion and inclusion criteria before considering it for discussion. Thus, Section 3.3 discusses the 12 selected papers. Table 4 presents the number of studies returned using the search string in each database, the total of studies removed using the criteria defined, and the papers selected as related work.

3.3 Discussion of Information Classification Studies

Table 5 summarizes characteristics extracted from the 12 studies selected. Security area refers to the goal of the application of NLP tasks. We identified studies evaluating different approaches to automate the information classification process (ALZHRANI et al., 2016b;

Table 5 – Characteristics extracted from state-of-the-art studies related to information classification.

Study	Security Area	Analysis Level	Multi-level Classification	Text Mining Tasks	Neural Network	PII entities
(ALZHRANI et al., 2016a)	DLP	Par	Yes	CL + CA	No	No
(ALZHRANI et al., 2016b)	IC	Par	Yes	TM + CA	No	No
(ALZHRANI et al., 2017)	DLP	Par	Yes	CA + CL + TM	No	No
(ALZHRANI et al., 2019)	IC	Par	Yes	CL	Yes	No
(GEETHA; KARTHIKA; KUMARAGURU, 2021)	PDD	Doc	No	CA	No	Yes
(HEINTZ et al., 2022)	IC	Doc+Meta	Yes	CA	Yes	No
(HUANG; CHIANG; CHANG, 2018)	DLP	Doc	Yes	CL + CA	Yes	No
(IBNUGRAHA; NUGROHO; SANTOSA, 2020)	RA	Doc	Yes	IR + IE	No	No
(PARK et al., 2016)	IC	Doc	Yes	IE + CA	No	Yes
(YANG et al., 2023)	PDD	Doc	No	CA	No	Yes
(YAZIDI et al., 2016)	DLP	Doc	No	CA	No	No
(ZARDARI; JUNG, 2016)	IC	Meta	No	CA	No	No
Proposed model	**IC**	**Doc**	**Yes**	**IE + CA + TM**	**No**	**Yes**

Security Area: (DLP) Data Leakage Prevention; (IC) Information Classification; (RA) Risk Analysis; (PDD) Personal Data Discovery.

Level Analysis: (Meta) MetaData; (Doc) All Document Content; (Par) Individual Paragraph.

NLP tasks: (CA) Text Classification; (CL) Text Clustering; (IE) Information Extraction; (IR) Information Retrieval; (TM) Topic Modelling.

Source: Created by the author.

ALZHRANI et al., 2019; HEINTZ et al., 2022; ZARDARI; JUNG, 2016; PARK et al., 2011). However, some studies targeted other security activities and used information classification to perform them. For example, studies proposing DLP systems used information classification to decide if an e-mail can be sent (HUANG; CHIANG; CHANG, 2018) or content can be uploaded to a cloud platform (ALZHRANI et al., 2016a). Information value is also relevant to performing a risk analysis because the value of information can be used to estimate security risks and define which should be prioritized (IBNUGRAHA; NUGROHO; SANTOSA, 2020). Authors also considered the presence of personal data for information classification (YANG et al., 2023; GEETHA; KARTHIKA; KUMARAGURU, 2021).

We also extracted the document content analyzed to classify information, identified as Analysis Level in Table 5. We identified one proposal using exclusively the document's metadata (ZARDARI; JUNG, 2016) and another combining the document's content and metadata (HEINTZ et al., 2022) to support information classification. A series of studies analyzed each paragraph individually and defined its classification level (ALZHRANI et al., 2016b; ALZHRANI et al., 2016a; ALZHRANI et al., 2017; ALZHRANI et al., 2019). The high granularity of this approach in information classification is very interesting to support access control methods since parts of the document can have access restrictions, whereas others can be available for access. The remaining studies considered the whole document to define the classification level. Document-level is the typical approach in the real-world scenario since it is not practical to manually perform security classification based on MetaData or for each paragraph in a lengthy document.

The studies also defined different approaches associated with classification levels. Some studies implemented a two-level information classification scheme that only differs documents containing sensitive information from those related to non-sensitive content (ZARDARI; JUNG, 2016; YAZIDI et al., 2016; YANG et al., 2023; GEETHA; KARTHIKA; KUMARAGURU, 2021). Although this approach can be helpful in specific situations, applying it to classify information in a real-world scenario is too simple because only one classification level does not provide sufficient information about sensitivity to implement appropriate security controls. Studies also proposed two different approaches associated with the multi-level classification.

The first multi-level classification approach considered static security levels based on an information classification scheme (ALZHRANI et al., 2016b; ALZHRANI et al., 2016a; ALZHRANI et al., 2017; ALZHRANI et al., 2019; HUANG; CHIANG; CHANG, 2018; HEINTZ et al., 2022). The approach analyzed each document and assigned one of the previously defined security levels. Real-world processes based on manual information classification usually apply this approach. The other approach used a score or the information value associated with confidential information in an electronic document (PARK et al., 2016; IBNUGRAHA; NUGROHO; SANTOSA, 2020). In this approach, the document is not directly classified, but the score or value is used as a reference to assign the classification level. This method offers a significant advantage over the previous approaches because it associates a quantitative value to the sensitivity level.

Another data extracted from related studies was the NLP tasks used to classify information. Following the pattern identified in the authors' SLR (IGNACZAK et al., 2021), 10 out of 12 approaches applied text classification, and five implemented other NLP tasks in sequence with text classification. Two studies integrated text clustering and text classification (ALZHRANI et al., 2016a; HUANG; CHIANG; CHANG, 2018). ALZHRANI et al. (ALZHRANI et al., 2016a) proposed a model that evaluates each document paragraph independently and groups similar paragraphs using clusters. In the proposed model, several clusters are generated, and the text classification task uses security features selected from each cluster. (HUANG; CHIANG; CHANG, 2018) proposed a system to classify e-mail messages using NN to perform text classification. The performance of NN demands a considerable amount of data to train the classifier, and the authors highlighted that imbalanced real-world datasets could impact NN's performance. Since a real-world dataset usually contains less confidential data than non-confidential data, the study proposed using text clustering to under-sample majority classes and make the amount of data similar to the minority class. The authors trained the NN classifier using the balanced dataset.

ALZHRANI et al. (ALZHRANI et al., 2016b) applied topic modeling and text classification. The study associated each paragraph of the documents in the training set with a sensitivity level and applied LDA to associate each paragraph with the main topics. The topics were used to remove the paragraphs that can negatively impact classifier performance. ALZHRANI et al. (ALZHRANI et al., 2017) analyzed the application of text classification, text clustering, and topic modeling. The authors proposed two approaches. The first approach extracted topics related to different sensitivity levels and performed a process to remove impure topics based on thresholds, so they trained a model using a Logistic Regression classifier. The second approach combined a text clustering approach using TF-IDF as a basis and a linear SVM classifier.

PARK et al. (PARK et al., 2016) proposed a system to measure data sensitivity, which uses multiple features to achieve a high-level classification accuracy. Among the features extracted before classification, the system identifies entities as company names and medical information using information extraction. IBNUGRAHA; NUGROHO; SANTOSA (IBNUGRAHA; NUGROHO; SANTOSA, 2020) proposed an adaptable classification process using information retrieval and information extraction. The study used keyword matching and regular expression to filter strings and patterns from a data source to perform information classification. The next step was the application of TF-IDF to calculate a weight to information and define the risk level.

We analyzed approaches applying NN to perform text classification and discovered three studies supported by different models. ALZHRANI et al. (ALZHRANI et al., 2019) implemented a Convolutional Neural Network (CNN) and word embedding and evaluated the performance using different datasets. The results demonstrated that the CNN model did not exceed a shallow classifier used as the baseline, and the authors justified that it occurred due to the use of small and imbalanced datasets in the evaluation. Huang et al. (HUANG; CHIANG; CHANG, 2018) evaluated the performance of NN for multi-level security classification, integrating text clustering and text classification. The study analyzed the classifier's performance based on true positive rates and evaluated the proposal in a real corporation. The performance of the NN classifier decreased in the real-world scenario, and the authors mentioned the diversity of content as the main reason. The study of HEINTZ et al. (HEINTZ et al., 2022) evaluates two pre-trained transformer models, DistilBERT and DistilRoBERTa, to classify information using a three-level information classification scheme. The study used a corpus based on a collection of documents from the Foreign Relations of the United States. The authors highlighted that the DistilRoBERTa model achieved highly accurate results if metadata and content were used as input.

We also evaluated which models considered PII entities in the information classification process. Laws and regulations regarding personal data protection stimulated studies to consider this kind of data. Yang et al. (YANG et al., 2023) proposed a compliance-driven security classification scheme guided by personal data laws and regulations. The scheme named Gen-DT combined Decision Tree and SVM to classify documents provided by two Chinese universities into regulated and non-regulated classes. The proposed scheme outperformed NN and shallow classifiers used as baselines. Geetha et al. (GEETHA; KARTHIKA; KUMARAGURU, 2021) proposed a system to classify social network posts as sensitive or not based on the presence of personal data. The system implemented a binary text classification and assessed the performance of three shallow classifiers in a corpus consisting of tweets. PARK et al. (PARK et al., 2016) implemented a system supported by rule-based approaches and machine learning methods to recognize 11 sensitive data types. As the study objective was to assign value to devices, the system assigned sensitivity scores to a device according to the sensitive data present in documents stored in it. Based on the sensitivity scores, the study estimated the device value. Interestingly, only PARK et al. (PARK et al., 2016) implemented information extraction to recognize PII entities; the others relied on text classification to identify documents containing personal data.

In a last analysis, we verified whether the studies addressed the challenge related to document reclassification and concluded that none of them focused on this issue directly. Although we understand it is not a model characteristic, studies could suggest approaches to dealing with this issue.

3.4 Discussion of NERC Studies

In Section 3.3, we observed that only three studies considered the presence of PII entities in the automation of the information classification process. As the model proposed in this study considers personal data for estimating the document value, we searched for proposals applying information extraction approaches to recognize PII entities. We observed that data discovery and privacy studies stand out in applying the task to identify PII. Several studies described rule-based assessments, but we selected only those that evaluated a machine-learning approach, including hybrid implementations combining rule-based and machine-learning algorithms. Table 6 summarizes the characteristics extracted from the eight studies associated with PII discovery.

Studies have applied machine learning classifiers to identify a set of coarse-grained entities. GENG et al. (GENG et al., 2008) implemented a hybrid approach to classify emails containing personal data. The study used a rule-based approach to identify four coarse-grained entities (email, phone number, address, and money) in the Enron dataset. The authors created a table using association rules based on the presence of the entities and implemented Decision Tree and SVM classifiers to predict if a message includes personal data.

Another study tested a NERC-based approach to de-identify PII and Protected Health

Table 6 – Characteristics extracted from state-of-the-art studies associated with PII discovery.

Study	Language	Number of Entities	Model(s)	Hybrid	Relation Extraction
(DIAS et al., 2020)	PT	18	CRF, RF, BiLSTM	Yes	No
(FENZ et al., 2014)	DE	18	SVM	Yes	Yes
(GENG et al., 2008)	EN	4	DT, SVM	Yes	No
(KAPLAN, 2020)	EN	6	BiLSTM-CRF	Yes	No
(LEITNER; REHM; MORENO-SCHNEIDER, 2019)	DE	19	BiLSTM-CRF, BiLSTM-CNN-CRF	No	No
(NAGPAL; DASGUPTA; GANESAN, 2022)	EN	134	NFGEC, BERTEC	No	No
(PEARSON; SELIYA; DAVE, 2021)	EN	5	Spacy e OpenNLP	No	No
(YAYIK et al., 2022)	TR	77	BiLSTM	Yes	Yes
Proposed model	**PT**	**23**	**BERT, BiLSTM-CRF**	**No**	**Yes**

Language: (DE) German; (EN) English; (PT) Portuguese; (TR) Turkish.
Models: (BiLSTM) Bidirection Long-Short Term Memory; (CNN) Convolutional Neural Networks; (CRF) Conditional Random Fields; (DT) Decision Tree; (NFGEC) Neural Fine Grained Entity Classification; (RF) Random Forest; (SVM) Support Vector Machine.

Source: Created by the author.

Information (PHI). It evaluated OpenNLP and Spacy libraries to recognize five entities (Name, Place, Address, Dates, and Numbers) in English medical documents (PEARSON; SELIYA; DAVE, 2021). KAPLAN (KAPLAN, 2020) analyzed the performance of Bi-LSTM-CRF architecture to recognize six personal data entities in call center transcriptions. The experiment used transcriptions of English-spoken calls taken by a call center and implemented a DL architecture using FLAIR. The study also evaluated the use of different embeddings in combination with FLAIR.

Although analyzing NERC performance in identifying the most common types of entities is very important, the current regulation scenario imposes the necessity of recognizing a more comprehensive number of personal data-related entities. Aiming to meet General Data Protection Regulation (GDPR) compliance, DASGUPTA et al. (DASGUPTA et al., 2018) proposed a set of 134 Personal Data Entity Types (PDET) and evaluated the performance of neural network-based models to recognize fine-grained entities in three English-related datasets. NAGPAL; DASGUPTA; GANESAN (NAGPAL; DASGUPTA; GANESAN, 2022) used the same set of 134 PDET to analyze the performance of BERTEC, a language model based on DistilBERT that incorporates embeddings as side information. The study evaluated BERTEC using three English datasets: OntoNotes, IMDB reviews, and Jigsaw Toxicity. The study of YAYIK et al. (YAYIK et al., 2022) proposed a solution based on a Bi-LSTM NN, rule-based model, and dictionary to extract 77 personal data entities from a dataset containing Turkish sentences. The study also proposed applying RE based on profiles using window sizes. The study considered that all entities inside a window size range belong to the same profile.

Two studies analyzed NERC approaches to recognize fine-grained entities in German datasets (FENZ et al., 2014; LEITNER; REHM; MORENO-SCHNEIDER, 2019). In the study of FENZ et al. (FENZ et al., 2014), the authors proposed a system to identify 18 types of patient data defined by the Health Insurance Portability and Accountability Act (HIPPA) in health records to document de-identification. The system implemented a hybrid approach, combining an SVM classifier with dictionaries and pattern matching. The system proposed RE handcrafted rules considering the entities' order in a document. LEITNER; REHM; MORENO-SCHNEIDER (LEITNER; REHM; MORENO-SCHNEIDER, 2019) implemented a NERC approach to recognize 19 fine-grained entities in legal documents. The authors created a dataset containing 750 German court decisions and evaluated the performance of six models comprising CRF and Bi-LSTM architectures.

The study of DIAS et al. (DIAS et al., 2020) focused on the Portuguese language and proposed a hybrid system for discovering 18 types of personal data, combining machine learning with rule-based and lexicon-based models. The personal information considered in the study refers to the Portugal data format. The study evaluated the performance of CRF, Random Forest, and Bi-LSTM classifiers in three datasets; one was created specifically for the assessment since public datasets contained only some searched entities.

3.5 Related Work Analysis and Research Opportunities

This section compares the proposed model to the information classification studies discussed in Section 3.3. As the model implements information extraction to discover PII entities, we also differ it from the studies presented in Section 3.4. Aiming to support the comparison of the proposed model to related work, we present the model characteristics at the bottom of the tables 5 and 6.

Most information classification studies implemented supervised approaches aiming at fitting a document into one of the sensitivity levels and applied a scheme based on two (GEETHA; KARTHIKA; KUMARAGURU, 2021; YAZIDI et al., 2016; ZARDARI; JUNG, 2016; YANG et al., 2023) or multiple levels (HUANG; CHIANG; CHANG, 2018; ALZHRANI et al., 2016b; ALZHRANI et al., 2016a; ALZHRANI et al., 2017; HEINTZ et al., 2022; ALZHRANI et al., 2019). One study proposed the association between a score and the document security classification and designed a method to estimate the score to apply in risk assessment (IBNUGRAHA; NUGROHO; SANTOSA, 2020). Thus, we identified an opportunity to propose a new approach to security classify a document based on its value. This approach was not found in the information classification domain, and it demonstrates this study's originality and represents the main difference between the proposal and related work.

This study proposes a model for estimating the information value of a document in monetary terms, which makes the model independent of the information classification scheme since it is not related to a specific number of sensitivity levels. PARK et al. (PARK et al., 2016) proposed a similar approach where the goal was not to perform the information classification but to estimate the value of an asset to identify the critical ones. Moreover, PARK et al. (PARK et al., 2016) did not detail the method used to estimate information value and only explained the NLP tasks at a high level. Unlike PARK et al. (PARK et al., 2016), this study estimates the value of each document, and it is directly related to the information classification research topic. The model proposed in this study can be associated with multi-level classification.

Two related studies applied topic modeling in the proposed information classification process, and both extracted topics to support the text classification task. This study differs in the application of topic modeling because it considers topic distribution to estimate the information value. This contrast results from the difference in the approaches implemented. Related work addressed the information classification as a classification problem while we handled it in a regression scenario. This distinction allows us to estimate the monetary worth of each document instead of its class. Based on our current knowledge, this is the first study to propose an NLP-based model to estimate the document value in the information classification field.

The proposed model considers personal data and document content in the information classification process, while related work adopts the approaches individually. Using personal data and information context is another opportunity in the research field since this integration offers more characteristics to classify documents correctly. The model proposes applying a NERC approach to identify PII entities that lead to information value estimation, as proposed by PARK et al. (PARK et al., 2016). We based our decision on the need to conform the information classification process to government laws and regulations, as introduced by ZARDARI; JUNG (ZARDARI; JUNG, 2016), YANG et al. (YANG et al., 2023), and GEETHA; KARTHIKA; KUMARAGURU (GEETHA; KARTHIKA; KUMARAGURU, 2021).

We also understand that the reduced number of languages associated with the corpora used in experiments that evaluate the automation of information classification is an opportunity. Analyzing information classification-related work, we identified one study using a Chinese corpus (YANG et al., 2023), and seven studies mentioned English corpora. Four studies did not introduce details about the corpora, but we can infer that they used English and Chinese content. So, we consider that the performance assessment of NLP tasks for automating information classification in other languages is a research opportunity, and this thesis explores it by evaluating the model using a Brazilian Portuguese corpus.

The NERC approach supporting the proposed model also comprises similarities and opportunities compared to related work. As implemented by YAYIK et al. (YAYIK et al., 2022), FENZ et al. (FENZ et al., 2014), LEITNER; REHM; MORENO-SCHNEIDER

(LEITNER; REHM; MORENO-SCHNEIDER, 2019), and DIAS et al. (DIAS et al., 2020), the proposed NERC approach targets a set of fine-grained personal data entities. However, this study focuses on PII that meets the Brazilian context and, consequently, is in accordance with Brazilian Law. The study of DIAS et al. (DIAS et al., 2020) implemented a system to discover entities in Portuguese texts, but it targeted entities related to documents issued by Portugal. In contrast, this study addresses entities related to Brazilian documents. Additionally, different from YAYIK et al. (YAYIK et al., 2022), FENZ et al. (FENZ et al., 2014), and DIAS et al. (DIAS et al., 2020), this study evaluates NERC approaches based exclusively on machine learning classifiers. We do not implement a hybrid approach to recognize entities.

Another aspect that differentiates this study from related work is the implementation of RE to filter entities that need to be protected. As mentioned by NAGPAL; DASGUPTA; GANESAN (NAGPAL; DASGUPTA; GANESAN, 2022), the presence of some entities in a text does not necessarily characterize PII, so as well as YAYIK et al. (YAYIK et al., 2022), we implemented a rule-based RE considering the distance between tokens to remove non-PII entities identified by the NERC approach.

4 PROPOSED MODEL

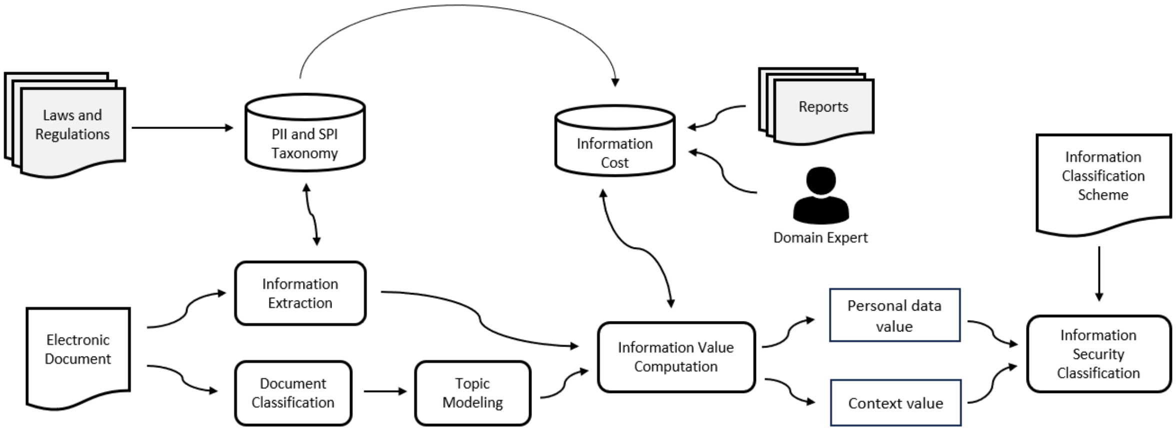
This chapter describes the proposed model to classify documents based on a value-based approach. Figure 4 presents the model's overview, highlighting the NLP tasks and external information necessary to estimate the information value. In the model description, we briefly present some decisions concerning the model implementation of the experiment. The decisions were based on regional information since the study was conducted in Brazil. Moreover, we mentioned algorithms and approach decisions because we implemented them in the experiments. It is necessary to highlight that we designed the model to offer adaptability, so implementations can use distinct algorithms from those employed to evaluate the model. Aiming for a better organization, we divided the model description into five sections as follows:

- Personal and Context Values: the model uses personal data and context values to estimate document worth. The estimation needs internal and external data as a reference to calculate each value, and we propose an approach based on reports and domain experts in Section 4.1.
- Entity Analysis: the model needs to recognize entities related to personal data, and we detail the use of NERC and RE in Section 4.2.
- Topic Analysis: we suggest implementing topic modeling to extract the main themes from documents with distinct values to use as the basis for context value estimation. We describe the approach suggested in Section 4.3.
- Information Value Computation: the model uses the entities, topics, and internal and external data to estimate personal and context values. We detail the strategy in Section 4.4.
- Information Security Classification: we present how we translate the information value to a sensitivity level in Section 4.5.

4.1 Personal and Context Values

The model uses two values for estimating the document's worth: personal data and organizational context values. The first refers to the value of a document containing PII and Sensitive Personal Information (SPI). In recent years, discussions about privacy have sparked laws and regulations to impose obligations on organizations about personal data collection and processing. For example, we can highlight GDPR, created by the European Union, and the California Consumer Privacy Act (CCPA), the California state law that ensures privacy rights to consumers. The federal government sanctioned LGPD in Brazil, which defines sensitive

Figure 4 – The proposed model to classify information based on the information value.



Source: Created by the author.

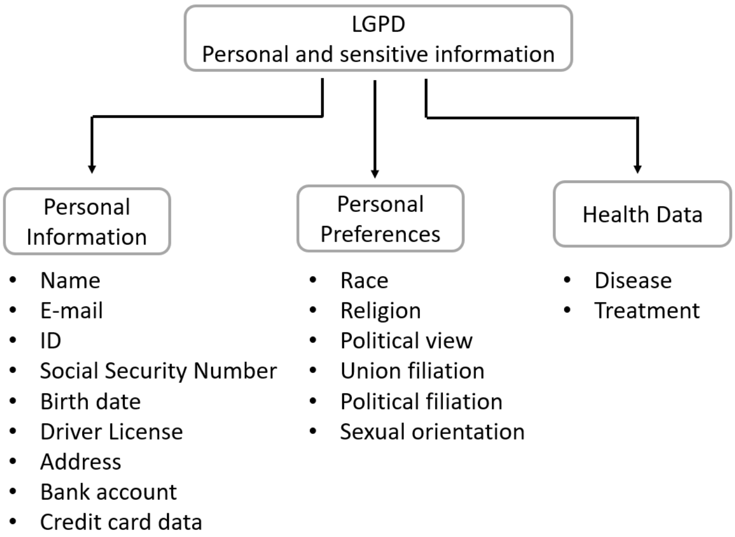
data associated with Brazilian citizens. Additionally, regulations can specify data protection for specific industries. One example is HIPAA, which aims to protect the privacy and security of health information.

The aforementioned laws and regulations established the possibility of penalties for organizations in case of information security incidents comprising PII and SPI. In the case of data breaches, costs for organizations can achieve hundreds of millions of dollars in fees and fines for failure to notify affected customers and governing agencies (POYRAZ et al., 2020). The result of this scenario is that personal data have become a valuable asset for organizations. Therefore, the model considers PII and SPI associated with laws and regulations included in documents to estimate the information value. In this study, the proposed model considers the Brazilian LGPD's PII and SPI since it is being carried out in Brazil. However, there is no restriction to relate other laws and regulations to the model.

An OECD study surveyed methodologies for measuring personal data's monetary value. Although it considered that no methodology is commonly accepted, it mentioned two possible approaches to use in the estimation: (i) market valuations of personal data or other related market measures and (ii) individual perceptions of the value of personal data and privacy (OECD, 2013). We identified scientific studies and corporate reports estimating personal data value based on analyzing individual perceptions (LI; LIU; MOTIWALLA, 2021; KALETA; MAHADEVAN; THACKSTON, 2023; ORANGE, 2014) and market information (IBM Security, 2023). In this study, we chose to evaluate the model using the market-information approach because we believe it is unaffected by the influence of user subjectivity, albeit this is not a model restriction.

We propose associating a taxonomy of PII and SPI with studies or reports so the model implementation can use this relation for estimating the document's personal data value. We recommend a taxonomy-like structure because it can link one or multiple PII with a value. Figure 5 presents a simple taxonomy based on the Brazilian LGPD law, highlighting Brazilian

Figure 5 – An example of taxonomy presenting PII and SPI according to Brazilian law.



Source: Created by the author.

citizen information that organizations should protect. In order to avoid redundancy, in the rest of the study, we refer to PII encompassing sensitive and non-sensitive personal information.

On the other hand, a document does not need to contain personal information to be considered valuable. Documents about intellectual property, strategic reports, and various others hold significant organizational value. So, the second value considers the context of information in the organization. Unlike personal data, there is no external guide to support information value estimation derived from its content; instead, the assessment relies on organizational knowledge. The research community has made many efforts to value information in the last 20 years, and it remains a challenge due to the intangible characteristics of information (FLECKENSTEIN; OBAIDI; TRYFONA, 2023). FLECKENSTEIN; OBAIDI; TRYFONA (FLECKENSTEIN; OBAIDI; TRYFONA, 2023)'s study surveyed existing approaches for data valuation and grouped them into market-based, economic, and dimensional models.

In this study, we selected the market-based approach to support the context value estimation. We utilized organizational data to assess the information value, considering the cost for the organization if the information is compromised in an information security incident. We also relied on domain experts (i.e., organization employees) to determine the organizational context value since they are the organization's knowledge custodians. It is essential to highlight that the model does not restrict the approach to establishing personal data and context values so that other strategies can be adopted.

4.2 Entity Analysis

In order to identify PII entities, we indicate applying the two information extraction sub-tasks: NER and RE. Identification of PII is the central factor in determining the value of personal data and contributes to the computation of context value alongside the topic analysis. In the first stage, NER is used to identify entities associated with personal data. Given the propensity for traditional rule-based methods, such as regular expressions, to generate a considerable number of false positives (SAHA et al., 2020), we recommend implementing a NERC approach to identify the entities.

Recognizing PII entities brings an additional challenge when compared with other entity types. The same entity type can be defined as personal or not since the decision depends on whether it is related to a person. For example, if a phone number is related to someone, it must be considered personal data. On the other hand, a phone number associated with an organization is not personal data. So, after recognizing the entities, it is relevant to identify relationships among them, considering that some entities similar to personal data in a document could not express sensitivity. The RE sub-task performs a primary role in overcoming the challenge of recognizing personal data entities. As supervised RE methods demand the need to annotate many training data and are hard to extend (SMIRNOVA; CUDRÉ-MAUROUX, 2018), we adopted the RE approach proposed by EL-ASSADY et al. (EL-ASSADY et al., 2017), which uses distance restriction to define the relationship.

4.3 Topic Analysis

Additionally to information extraction, the model defines a second approach to identifying the information value, which comprehends two activities. The first activity, named document classification, predicts the organization department associated with a document. We consider this association relevant because each department can use specific terms to produce content, so document classification applies the text classification task to predict the department linked with the content. Based on the initial tests performed by the authors, whose results are presented in Chapter 6, we identified that the performance improved when training one model for each department, so document classification is an essential activity to allow this approach. The model implemented the BoW, applying the Term Frequency-Inverse Document Frequency (TF-IDF) for classifying documents. TF-IDF is an approach to establishing a weight for each word based on the number of occurrences in a specific document and the number of documents comprising it (KOWSARI et al., 2019).

After text classification, the model applies topic modeling to extract document context. Topic modeling uses unsupervised machine learning techniques to identify the most prevalent themes in a corpus, expressed in sets of related words (CHURCHILL; SINGH, 2022). As the model aims to estimate the value of information, topic modeling needs to extract a number representing the most prevalent topics. In order to achieve this, the model extracts the topic distribution from the document using a set of topics previously identified as a reference. Thus, the proposed model needs a training corpus to generate a model for each department.

We decided to implement the LDA model because it is the most commonly used model to identify topics in text (ZHENG et al., 2023). LDA is a generative statistical model that assigns the probability or weight of each topic in a document, assuming that a topic represents a distribution of words (HAPKE; LANE; HOWARD, 2019). A drawback of the LDA model is the need to specify the number of topics to be extracted as a model parameter. In the experiment, we implemented the approach proposed by ZHAO et al. (ZHAO et al., 2015) that defined a rate based on perplexity named Rate of Perplexity Change (RPC).

4.4 Information Value Computation

We base the computation of the information value on three inputs: (i) the entities recognized in the document, (ii) the distribution of the topics in a document considering its department, and (iii) the costs related to personal data and context information. Using these inputs, the model estimates the personal data and context values.

In order to estimate the personal data value, the model receives the cost of personal data based on a previous assessment. Two approaches that can be applied are mentioned in Section 4.1. The activity considers that every PII (or a group of PII) in the taxonomy has a cost. The model counts the number of each type of PII entity discovered in the document and multiplies them by the respective cost. The model estimates the personal data value via a straightforward calculation. The estimation of personal data value considers only PII entities identified in a document, not using topic distribution.

The organizational context value assessment demands a more complex calculation, and we recommend implementing a regression model. A regression is an equation representing how a set of factors explains an outcome variable and how each factor impacts the outcome (ARKES, 2023). Several regression models differ in how the outcome variable is estimated (GRÖMPING, 2015) and, in machine learning, regression methods are designed to predict continuous numeric outputs based on input variables relations (FERNÁNDEZ-DELGADO et al., 2019). It is essential to emphasize that the proposed model is not strictly associated with any regression model. Furthermore, the proposed model can use other approaches than a regression model, supervised or not.

For estimating the context value, the model receives the numeric feature set extracted from information extraction and topic modeling tasks as input. The topic modeling task produces the probabilities associating each topic with a document, and the model uses them as topic features. The total of topic features is variable since it depends on the number of topics extracted from a document. Although the topics represent the main themes associated with a document, it is important to consider personal data features to support the regression model. The output of the

Table 7 – An example of a three-level information classification scheme to assign the sensitivity label.

Label	Lower Bound	Upper Bound
Public	-	10.0
Internal	10.01	300.0
Confidential	300.01	-

Source: Created by the author.

information extraction task is a list of PII entities discovered in the document. We understand that the frequency of PII entities does not necessarily impact the context value, but its presence does. So, the model considers a set of personal data features representing the presence of groups of entities in a document.

As the model implements a regression, a type of supervised learning, it is necessary to train a model. We extracted a complete feature set from each corpus document for the training phase and associated it with the respective context value. As mentioned in Section 4.1, the context values of the set of documents comprising the training corpus are assigned by domain experts. The regression-trained model is used to estimate the context value of new documents.

4.5 Information Classification

The last stage is to label the document with its sensitivity level. Initially, the model adds personal data and context values to get the document value. In order to assign a label to the document, the model uses an information classification scheme that relates each sensitivity level to a range of values. We considered that the more valuable a document is, the higher its sensitivity level. The model utilizes an information classification scheme to assign labels by referencing the appropriate range of values determined by the computed value of the document. Table 7 depicts an example of a three-level information classification scheme. In the example, if the estimated value for a document is lower or equal to $ 10.0, it is classified as "Public"; if the estimated document value is higher than $ 10.0 and lower or equal to $ 300.0, it is assigned with an "internal" label; and a document with the value higher than $ 300.0 is classified as "Confidential".

5 MATERIALS AND METHODS

This chapter presents the methodology to implement and evaluate the proposed model, aiming to answer the research question. This work employs the experimental design method. According to EDGAR; MANZ (EDGAR; MANZ, 2017), "experimentation is good for testing theoretical models". The authors also declared that applied research is essential in cybersecurity because it allows the comprehension of how well the knowledge produces systems to solve problems and generate predictable outcomes. So, applied experimental methods provide the understanding of a system's performance or effectiveness (EDGAR; MANZ, 2017).

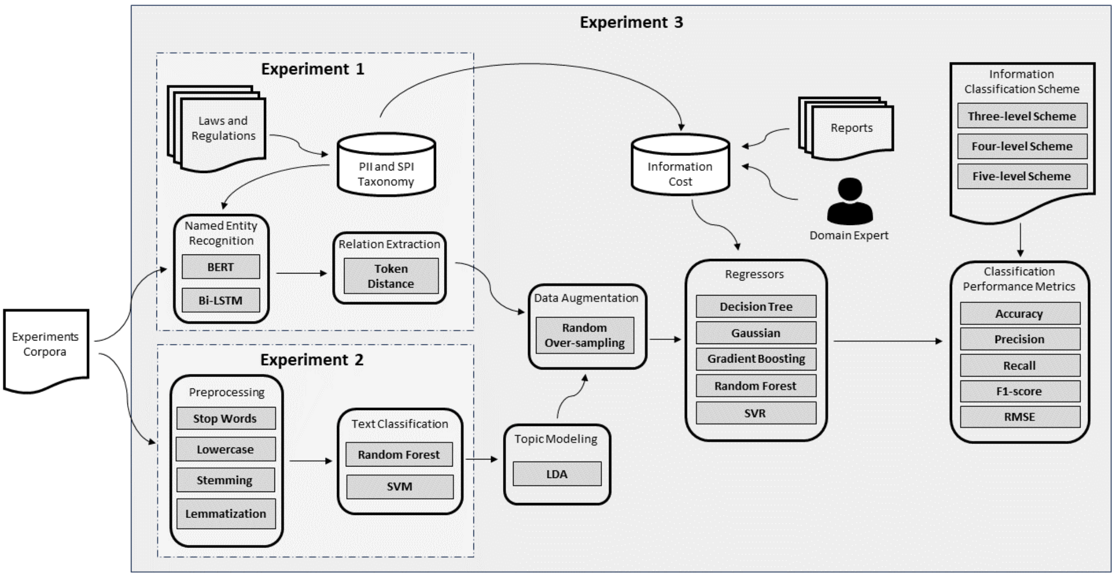
A relevant aspect of applied research is dealing with real-world challenges because it can address society's concerns. Thus, EDGAR; MANZ (EDGAR; MANZ, 2017) declares that "applied experimental methods leverage controlled tests that attempt to capture real-world behavior to determine how the applied system behaves". In order to perform controlled tests, the experiment must run in an environment that offers monitoring and rigorous control. Furthermore, determining the system behavior demands the definition of metrics for quantitative performance evaluation. Although the experiments designed for this study have their reproducibility compromised since they rely on internal documents that are not publicly accessible, we describe the decisions made to provide solid foundations in order to ensure the validity of the results. Thus, we base overall decisions on the machine learning experiment design proposed by ALPAYDIN (ALPAYDIN, 2014).

The proposed model involves the application of different NLP tasks in a row, so we designed three experiments to evaluate it. Figure 6 presents an overview of the experiments. The first experiment, described in Section 5.2, evaluates the performance of NERC models and an RE approach to identify PII and sensitive information based on two corpora. The evaluation of NERC and RE was necessary to analyze whether the proposed model could rely on these sub-tasks to identify personal information in documents to estimate its value. The second experiment assesses the traditional BoW approach to classify documents into distinct departments. We explain in Section 5.4 that we trained different topic models for each department. Hence, we needed to confirm the performance of the text classification task in forwarding the documents to the correct model. Thus, we evaluated distinct combinations of preprocessing techniques and two shallow classifiers. Section 5.3 details the document classification experiment.

The third experiment assesses the whole model, employing trained models that achieved significant performance in experiments #1 and #2 and implementing topic modeling and regression. We implemented a data augmentation approach to balance the dataset before training five regressors. The experiment outcome was the classification assigned for each document, which makes it possible to evaluate the performance of the proposed model. Section 5.4 delineates the third experiment. Before describing the experiments, Section 5.1

presents the corpora utilized to evaluate the proposed model.

Figure 6 – An overview of the experiments designed for evaluating the proposed model.



Source: Created by the author.

We decided to use Python language to implement the experiments. Although SOUZA et al. (SOUZA et al., 2018) pointed out that the most used programming language for evaluating text mining tasks with Portuguese content is Java, we believe that Python offers libraries and resources to implement the tasks applied in the experiments.

5.1 Experiments Corpora

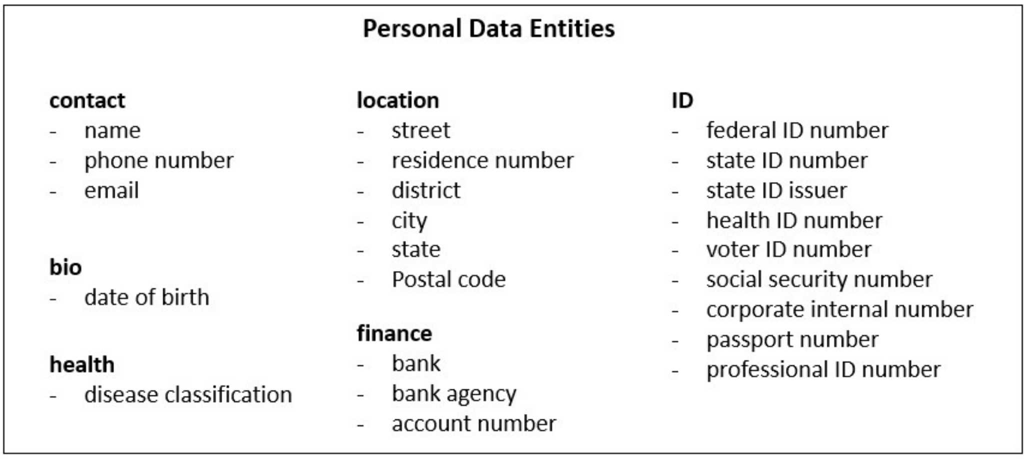
Throughout this study, we conducted distinct experiments utilizing different corpora tailored to specific research objectives and methodologies. The experiments' corpora are formed exclusively by documents written in Brazilian Portuguese. In the first experiment, we evaluated the capability of NERC approaches and RE to generalize and recognize PII entities in documents from distinct domains. So, we selected documents from two organizations representing different industry sectors. Due to the evaluation approach, we needed to manually verify the information extraction performance, so we selected a small number of documents from each organization. Section 5.1.1 described the corpus used in the first experiment. The remaining experiments needed a more robust corpus to assess the performance of the proposed information classification model, as well as the text classification task. Thus, we evaluated the second and third experiments using more documents from just one organization. Section 5.1.2 presents the corpus organized for these experiments.

5.1.1 Corpora and PII Entities for Experiment 1

The first experiment evaluated the performance of distinct NERC approaches and RE to recognize PII entities in documents associated with two different industry sectors. The proposed model uses the information extraction sub-tasks to identify PII compliant with LGPD. NERC is based on supervised learning, so we must annotate a document set. Thus, we defined the creation of an agnostic training corpus comprising documents from several industries, excluding those that provided documents for the testing corpus. This strategy also aimed to avoid bias in the trained models. Organizations comprising the education and health sectors provided real documents for establishing a testing corpus. Thus, the experiment relied on two distinct corpora, one used exclusively for training the NERC models and establishing the rules for the RE approach and a second corpus containing documents provided by the organizations and used only for testing purposes.

Before organizing the training corpus, defining the set of PII entities targeted in the first experiment was crucial. LGPD uses generic definitions and does not define which information should be considered personally identifiable. Therefore, we built the entity set based on the PDET taxonomy proposed by Dasgupta et al. (DASGUPTA et al., 2018). We also used a NIST guide for double-checking (MCCALLISTER, 2010). In order to define the entity set, we analyzed documents provided by the two organizations and compared the entities identified with those present in the taxonomy and the NIST guide. The result was a set of 23 personal data tags identifying Brazilian PII, presented in Figure 7.

Figure 7 – The set of 23 Brazilian-related PII evaluated in this study.



Personal Data Entities

contact
- name
- phone number
- email

bio
- date of birth

health
- disease classification

location
- street
- residence number
- district
- city
- state
- Postal code

finance
- bank
- bank agency
- account number

ID
- federal ID number
- state ID number
- state ID issuer
- health ID number
- voter ID number
- social security number
- corporate internal number
- passport number
- professional ID number

Source: Created by the author.

Additionally, we needed to add a 24th tag because many Brazilian documents contain the name of the city and the state written as a single word using a marked character (e.g., a hyphen) to segment, and the tokenizer function does not split into two tokens. So, we identified a total of 24 tags. It is worth noticing that the final set of documents selected for the testing corpus did not contain four existing entities shown in Figure 7: voter ID number, bank, bank agency, and account number. So, although we trained the models to consider 24 entities, the experiment recognized 20 of them.

We applied two approaches based on publicly available information to establish the training corpus: (i) we downloaded documents written in a natural language containing PII; (ii) we downloaded templates of documents requesting PII data and filled them with personal data. In order to implement the second approach, we searched the internet for files containing lists of PII data. The searches returned many documents comprising publicly available personal data, primarily on governmental sites. Additionally, we also found samples of leaked databases available on the internet. So, we merged the PII data collected with the templates to create documents written in natural language encompassing PII. The two approaches produced 1.109 documents.

We analyzed the training corpus and verified that some number-based PII entities were found in different formats. Since some formats are less frequent than others, we decided to implement a data augmentation approach. Data augmentation encompasses strategies for increasing the number of texts in a training corpus through diverse new textual content (FENG et al., 2021a). Data augmentation constructs synthetic new or modified texts based on existing content, increasing the size and enhancing the training corpus's quality to improve the machine learning model (SHORTEN; KHOSHGOFTAAR; FURHT, 2021; SHORTEN; KHOSHGOFTAAR, 2019). The application of data augmentation is valuable in scenarios where the volume of data is restricted, as usually occurs in those comprising personal data.

Data augmentation performed in this study aimed to increase the number of entities found in less frequent formats, and we focused on entities related to ID numbers, such as federal ID and state ID. So, we implemented scripts to create copies of some documents included in the training corpus and, in the new document, replace the format of the ID numbers for those less frequent. We created a list of possible formats for each ID-related entity and implemented a pseudorandom function to select one of them. The script generated 122 new documents containing the target entities in different formats. As the entity length could differ depending on the format picked by the scripts, the new documents needed to be tagged.

We also implemented a second data augmentation approach intending to expand the corpus by modifying the texts of the documents. We evaluated using Brazilian word embeddings to replace the nouns in the original documents for others that fit in the context. For this, we implemented GloVe embeddings trained with Brazilian documents available publicly[1]. However, our attempt did not succeed because noun substitution generated new documents that did not correspond to the context of real-world documents, so we discarded the produced documents. Therefore, the final training corpus comprised 1,231 documents, involving the initial 1,109 documents plus the 122 generated in the first data augmentation approach.

In order to evaluate the trained models' performance, we selected 40 distinct documents containing personal and sensitive data from each organization, so the testing corpus consisted

[1] http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc

of 80 documents. We took care to choose documents with different structures. The health organization provided medical and administrative documents, such as health records, exam analysis, and terms of donations. The educational organization supplied academic and administrative documents such as academic contracts, declarations, and scholarship terms.

5.1.2 Corpus for Experiments 2 and 3

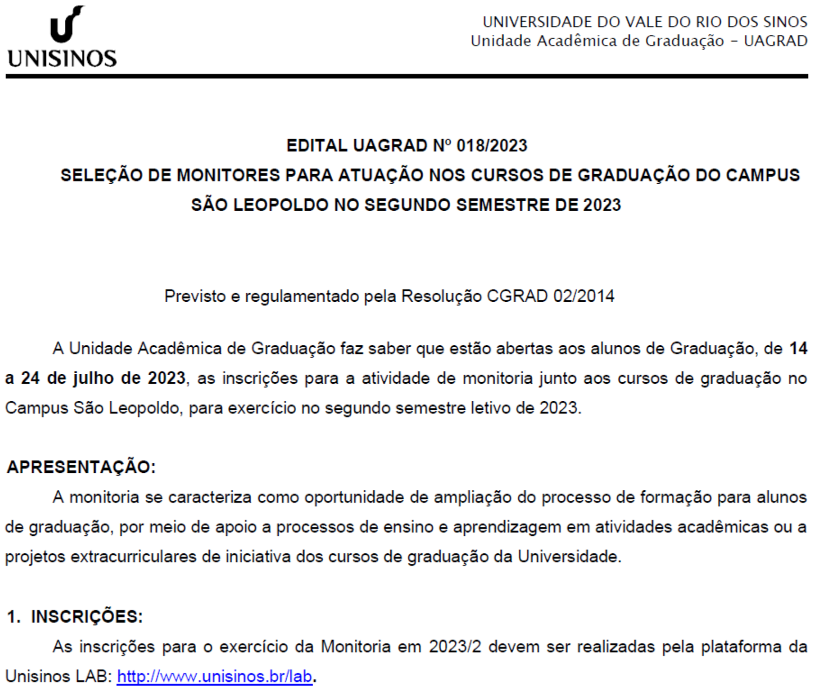
For the remaining experiments, we selected a corpus comprising 197 real-world documents provided by four different departments of the educational organization. We were careful in the corpus selection to ensure that its diversity did not bias the experiment's result, so the corpus comprehends documents with some similarity in the text structure and others completely different from each other. We also took care to select a very diverse corpus consisting of documents associated with distinct sensitivity levels. Finally, we included documents containing personal data and others without this data type.

To illustrate the heterogeneity of the corpus, we analyzed various characteristics, including the number of tokens and sentences across the 197 documents. We observed a wide range of sentence counts, with the lowest being one. Two documents contained only one sentence each. Conversely, one document contained 378 sentences. The average number of sentences across all documents was 24.76, with a median of 10. With respect to the number of tokens, document lengths varied significantly, with the smallest document comprising 50 tokens and the largest containing 11,237 tokens. On average, documents contained 905.59 tokens, with a median of 434.

The following departments provided documents for the experiments: marketing, academic secretariat, document management, and diploma issuance. The marketing department provided several documents containing information to publicize the educational organization. Marketing documents are entirely different from each other and have no similarity in their structure. The academic secretariat provided administrative documents such as internal regulations, academic reports, and public announcements. Most academic secretariat documents exhibit a completely different structure and content, with only a few sharing a similar format. Figure 8 presents a public announcement provided by the academic secretariat. This is an example of a document that holds some similarity with others since some sections can have analogous content.

The document management department is responsible for issuing and storing students' documents. The department provided academic documents such as undergraduate contracts, declarations, and scholarship terms. A sample of the documents provided by this department was used in the first experiment. The last department, diploma issuance, is in charge of emitting documents related to the conclusion of the undergraduate course. Unlike marketing and academic secretariat, the document management and diploma issuance departments are part of one of the biggest divisions of the educational organization and, therefore, manage documents with a few similarities. We believe that this situation occurs in many organizations,

Figure 8 – Example of a four-page public announcement provided by the academic secretariat.



UNISINOS

UNIVERSIDADE DO VALE DO RIO DOS SINOS
Unidade Acadêmica de Graduação – UAGRAD

EDITAL UAGRAD Nº 018/2023
SELEÇÃO DE MONITORES PARA ATUAÇÃO NOS CURSOS DE GRADUAÇÃO DO CAMPUS SÃO LEOPOLDO NO SEGUNDO SEMESTRE DE 2023

Previsto e regulamentado pela Resolução CGRAD 02/2014

A Unidade Acadêmica de Graduação faz saber que estão abertas aos alunos de Graduação, de **14 a 24 de julho de 2023**, as inscrições para a atividade de monitoria junto aos cursos de graduação no Campus São Leopoldo, para exercício no segundo semestre letivo de 2023.

APRESENTAÇÃO:

A monitoria se caracteriza como oportunidade de ampliação do processo de formação para alunos de graduação, por meio de apoio a processos de ensino e aprendizagem em atividades acadêmicas ou a projetos extracurriculares de iniciativa dos cursos de graduação da Universidade.

1. INSCRIÇÕES:

As inscrições para o exercício da Monitoria em 2023/2 devem ser realizadas pela plataforma da Unisinos LAB: http://www.unisinos.br/lab.

so including these departments is essential to establish a real-world-like corpus.

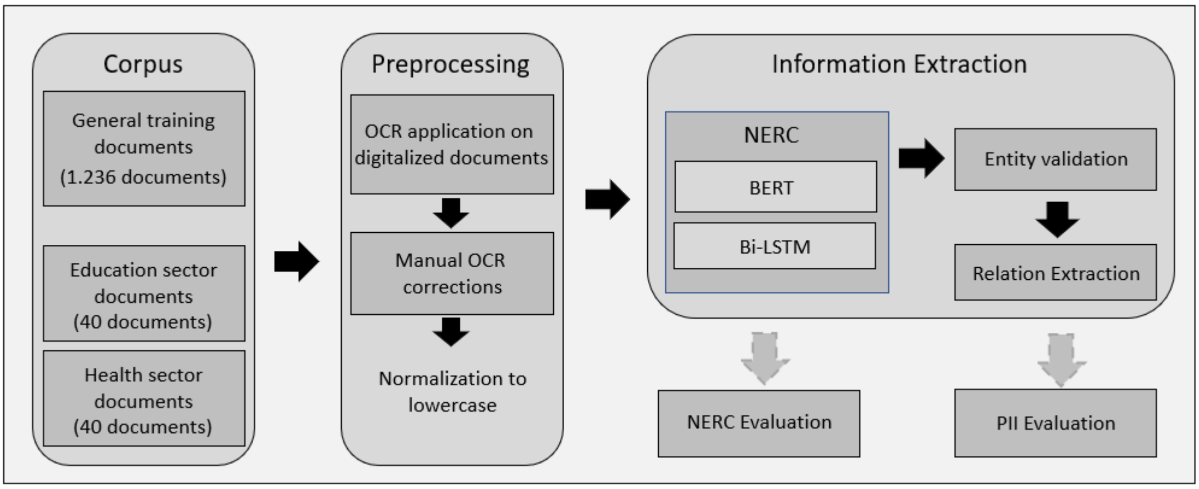
5.2 Experiment 1: PII Discovery Evaluation

This section details the experiment designed to evaluate the different machine-learning approaches to discover PII entities, which follows the process presented in Figure 9. In the first stage of the process, we annotated the training corpus described in Section 5.1.1. The following stage comprised preprocessing training and testing corpora. We set distinct versions of the training corpus and employed OCR-based scripts to convert digitized documents into text format within the testing corpus. In the last stage, we performed two evaluations regarding the PII discovery: (i) the performance of the NERC approaches before RE in identifying personal data and non-personal data entities contained in the documents, and (ii) the performance after RE considering only PII entities. The following subsections detail the experiment design and settings.

5.2.1 Annotated Entities

NERC approaches depend on tagged documents, so we needed to annotate the 1,231 training corpus documents. We used Doccano as the annotation platform to label the entities (NAKAYAMA et al., 2018). As we tagged documents containing personal and sensitive data, we opted for a tool allowing local installation. Two annotators manually tagged the training

Figure 9 – The PII discovery process implemented in the experiment.


Corpus
General training documents (1.236 documents)
Education sector documents (40 documents)
Health sector documents (40 documents)
Preprocessing
OCR application on digitalized documents
Manual OCR corrections
Normalization to lowercase
Information Extraction
NERC
BERT
Bi-LSTM
Entity validation
Relation Extraction
NERC Evaluation
PII Evaluation

Source: Created by the author.

corpus. The primary annotator was the author, assisted by a professional hired for the task. In order to ensure that annotations follow the same pattern, all tags were reviewed by the primary annotator. The total of annotated tags was 16.551. Although it is not possible to ensure the balance of the tags due to the documents' structure, we defined a minimum of a hundred occurrences for each annotated tag.

5.2.2 Preprocessing

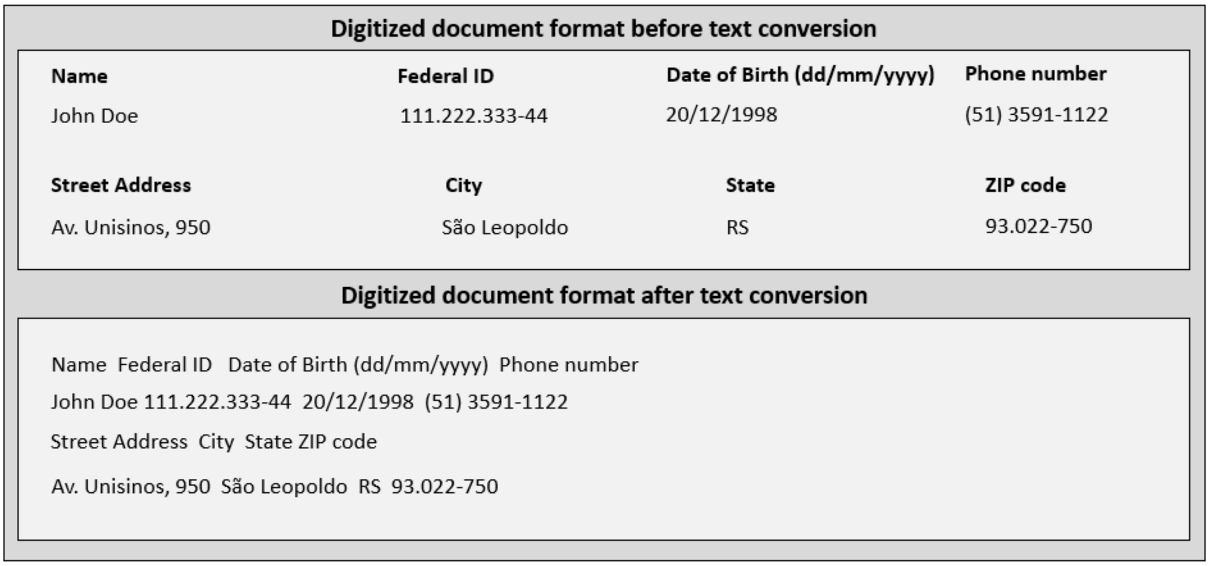
The preprocessing stage comprises the application of techniques in the raw text, improving the quality of the unstructured data processed by an NLP task (UYSAL; GUNAL, 2014). Applying preprocessing techniques aims to enhance pattern extraction and improve the performance of an NLP task. In this experiment, we employed different preprocessing techniques in the training and testing corpora. The training corpus involved documents in PDF, DOC, and DOCX formats, so we initially converted all to plain text. As the study of CASTRO; SILVA; SOARES (CASTRO; SILVA; SOARES, 2018) concluded that normalization was the preprocessing technique that produces the most significant improvement in the performance of a NERC experiment, we decided to evaluate the impact of normalizing the text to lowercase for training the models. So, we kept one training corpus containing each document in the original text format and created a second training corpus with all texts converted to lowercase. Consequently, we trained two models for each evaluated approach, the first comprising the original texts and named as cased, and the normalized one named uncased.

We also implemented preprocessing techniques in the testing corpus. The two organizations delivered PDF documents, and we converted all of them to plain text format. Some PDF documents were generated from a digitization process, so we implemented a script to perform OCR and convert the documents to plain text. The script utilized the open-source

OCR engine Tesseract (SMITH, 2007). In some cases, the OCR application resulted in errors in the plain text, and we manually corrected them. Although these errors will occur in real-world applications, we believe procedures can be implemented to correct most of them automatically. The most common error perceived was the misconversion of electronic addresses.

As some PDFs result from the digitization of semi-structured physical forms (i.e., medical records), the conversion of these documents resulted, in many cases, in the name of the form fields in one line and the values associated with them in a line below. We pondered about removing these documents from the testing corpus; however, we decided to keep them because they were examples of real-world documents, and we wanted to evaluate the models for real-world application. Figure 10 presents an example of a digitized form-based document in the original format and after its conversion to plain text. Another conversion issue related to this document type is that the conversion resulted in short paragraphs. We dealt with these cases by merging paragraphs containing less than 40 characters with the next paragraph.

Figure 10 – Format of files resulting from digitized forms before and after applying OCR and plain text conversion.



Digitized document format before text conversion

Name	**Federal ID**	**Date of Birth (dd/mm/yyyy)**	**Phone number**
John Doe	111.222.333-44	20/12/1998	(51) 3591-1122
Street Address	**City**	**State**	**ZIP code**
Av. Unisinos, 950	São Leopoldo	RS	93.022-750

Digitized document format after text conversion

Name Federal ID Date of Birth (dd/mm/yyyy) Phone number
John Doe 111.222.333-44 20/12/1998 (51) 3591-1122
Street Address City State ZIP code
Av. Unisinos, 950 São Leopoldo RS 93.022-750

Source: Created by the author.

5.2.3 Selection of NERC Approaches

In order to advance in selecting state-of-the-art algorithms for personal information discovery, we conducted research into recent surveys that provide comprehensive insights into the application of shallow and DL approaches for recognizing entities (LI et al., 2020; EHRMANN et al., 2023). EHRMANN et al. (EHRMANN et al., 2023) pointed out CRF as the leading shallow classifier implemented in the NERC task. When considering DL approaches, LI et al. (LI et al., 2020) and EHRMANN et al. (EHRMANN et al., 2023) emphasized the implementation of LSTM and Transformers for entity recognition. We also analyzed a survey comprising the application of NERC in the cybersecurity domain, and CRF, LSTM, and Transformers approaches were applied for recognizing entities in different cybersecurity activities (GAO et al., 2021).

During the design of the experiment, we considered implementing the CRF classifier. The study of AMARAL; VIEIRA (AMARAL; VIEIRA, 2014) demonstrated that a system based on the classifier overcame other systems to identify entities in Brazilian Portuguese texts. We implemented CRF using the Spacy-CRFsuite [2], but the model performed poorly in some tests, and we excluded it from the experiment. Due to removing CRF, the experiment evaluated only DL approaches. The first approach employed LSTM, more precisely, the birectional LSTM+CRF combination proposed by HUANG; XU; YU (HUANG; XU; YU, 2015). We chose the combination because it achieved the best overall performance in recognizing entities in a NERC contest comprising Brazilian Portuguese texts (COLLOVINI et al., 2019).

In addition, the study of AKBIK; BLYTHE; VOLLGRAF (AKBIK; BLYTHE; VOLLGRAF, 2018) demonstrated that the use of stacked embeddings improved the performance of Bi-LSTM+CRF approach in the NERC task, so we stacked Portuguese-trained Flair's forward and backward contextual embeddings with GloVe word embedding as input for our models. The GloVe vector dimension utilized was 300. The implementation used the FLAIR framework (AKBIK et al., 2019), and we set the number of hidden layers to 256, the number of epochs to 100, and the learning rate to 0.1.

The Transformer architecture also achieved significant results in performing NERC on Brazilian Portuguese texts (SOUZA; NOGUEIRA; LOTUFO, 2019). For implementing the Transformer approach, we opted for fine-tuning the BERTimbau[3] model because it overcame other models in a NERC experiment (SOUZA; NOGUEIRA; LOTUFO, 2020). In order to fine-tune the model, we set the learning rate to 0.0001, the batch size to 16, and three epochs.

Unlike occurs for the English Language, Portuguese is considered a low-resource language (NETO; SILVA; SOARES, 2023). Consequently, there is a lack of resources regarding the NERC application for the Portuguese language (ROCHA et al., 2022). To the best of our knowledge, this experiment pioneers the discovery of Brazilian PII, so we could not find other machine learning models to compare with the performance obtained by the models trained. Due to this, we did not establish baselines and focused on comparing the models trained from the corpus crafted for this study.

5.2.4 Post-processing and Relation Extraction

Before applying the RE task, we performed validations regarding some types of entities. We checked all entities recognized as the tag city against a dictionary containing all Brazilian

[2] <https://github.com/talmago/spacy_crfsuite>

[3] https://huggingface.co/neuralmind/bert-base-portuguese-cased

cities. We applied Levenshtein distance in the comparison to handle typos and abbreviations in the documents. As many entities refer to Brazilian ID numbers that follow a strict number of digits, we validated the entities' length to mitigate the chances of identifying a wrong value as an entity.

RE performed a primary role in the scope of this study because we targeted the recognition of PII entities related to Brazilian data protection law. We considered RE a relevant task to reduce the number of false positives on PII discovery because it allows the recognition of entities associated with an individual. Due to this context, our proposal identifies a particular entity and recognizes if it is related to someone. For example, a phone number can be associated with a person or a company, and according to Brazilian LGPD law, it must be protected only in the first association.

We employed the RE method proposed by EL-ASSADY et al. (EL-ASSADY et al., 2017), which uses distance restriction to define the relationship. The method defines that two or more entities must be close to each other (i.e., in the same paragraph) in the document to be considered valuable. We implemented an approach similar to YAYIK et al. (YAYIK et al., 2022) that defined a window to establish an entity as PII. The window consists of the number of existing tokens between two tokens in the text. However, unlike YAYIK et al. (YAYIK et al., 2022), who evaluated the window distance between predefined keywords and an entity, our approach was based primarily on the distance between two recognized entities. We applied the method and defined the token distances based on a previous analysis comprising training corpus documents.

5.2.5 Experiment Evaluation

We evaluated the experiment in two stages of the process. Initially, we assessed the result of the NER task, in which we considered the approaches' performances in identifying the entities tagged in the training corpus. In this stage, we did not consider if a system must protect the recognized entity. For example, identifying a company's phone in a document was considered correct in this stage because the phone was one type of entity tagged in the training data, even if the entity was not related to PII. In the second stage, we evaluated the performance after applying the RE task. In this assessment, we considered as an entity only those that an organization must protect. So, the recognition of a company's phone mentioned in the previous example is considered incorrect at this stage, and we focused only on PII-related entities.

The metrics used for experiment evaluation were Precision, Recall, and Fscore. We adopted the sets of metrics defined in SEGURA-BEDMAR; FERNÁNDEZ; ZAZO (SEGURA-BEDMAR; FERNÁNDEZ; ZAZO, 2013) because they allowed us to analyze if the approaches were identifying tags' boundaries or assigning the tags incorrectly. SEGURA-BEDMAR; FERNÁNDEZ; ZAZO (SEGURA-BEDMAR; FERNÁNDEZ; ZAZO, 2013) presented four approaches to calculating the metrics: Strict, Exact, Partial, and Type.

The Strict approach considers that the boundaries of an identified entity must comprise only the correct tokens and the tag must be appropriately assigned. The Exact type only cares about the tag boundaries, so an entity must be considered in cases where the tag is assigned incorrectly as long as the tokens were identified correctly. The Partial type considers only the partial matching of the entity boundaries and does not consider the assigned tag type. Finally, the Type approach only verifies if the type of an entity was assigned correctly and does not check if the boundaries match (SEGURA-BEDMAR; FERNÁNDEZ; ZAZO, 2013).

Considering that the objective of the experiment was to analyze the use of NERC and RE to identify personal data in documents, we believe that evaluating the performance of the models only using the strict approach is not appropriate because, in real-world applications, the recognition of partial entities is enough to discover documents that need to be protected. Therefore, to measure the four approaches, we manually extracted from each document the entities that should be recognized by NERC and which of these entities are related to PII and should remain after the RE stage.

5.3 Experiment 2: Document Classification

The proposed scheme applies topic modeling to identify the main topics related to confidential and non-confidential documents. As described in Section 4.3, the first activity of topic analysis involves categorizing documents conforming to the organization department, and this experiment assessed the text classification performance in this activity. The experiment evaluated the preprocessing techniques and machine learning classifiers.

The first evaluation was the impact of preprocessing techniques on text classification performance. Different studies demonstrated that distinct techniques had different impacts on the results of NLP tasks (ETAIWI; NAYMAT, 2017; UYSAL; GUNAL, 2014; KUMAR; HARISH, 2018). Based on UYSAL; GUNAL (UYSAL; GUNAL, 2014), we decided to use the following preprocessing techniques: stop word removal, lowercase conversion, and stemming. The authors evaluated the techniques in the study and concluded that the "appropriate combinations of preprocessing tasks depend on the domain and language". So, we designed an experiment to evaluate the impact of preprocessing techniques using the most complete corpus, the educational one. The health organization has not provided enough documents from different departments to allow us to implement a solid experiment.

We did not include the PoS technique because it did not achieve the best results in the evaluation performed by OLIVEIRA; MERSCHMANN (OLIVEIRA; MERSCHMANN, 2021). However, we included the lemmatization technique, explained by SAMMUT; WEBB (SAMMUT; WEBB, 2017) as a more sophisticated word transformation that replaces a word with its normalized form. The authors also described lemmatization as a valuable technique in languages with many different forms of the same word. Portuguese is an example of a language with this characteristic. Studies evaluating the impact of preprocessing techniques mentioned the interest in testing lemmatization in future work (KUMAR; HARISH, 2018; HACOHEN-KERNER; MILLER; YIGAL, 2020), and a study that included the technique in the evaluation presented that lemmatization impacted positively in the performance metrics (MALI; ATIQUE, 2021).

We opted to implement the BoW approach and the TF-IDF term weighting method. Before deciding on evaluating only the traditional TF-IDF, we analyzed approaches combining TF-IDF and static word embedding (BIRUNDA; DEVI, 2021). However, an analysis of SOUZA; FILHO (SOUZA; FILHO, 2023) concluded that the traditional TF-IDF outcome its combination with word embedding.

We implemented tokenization, stop word removal, and stemming using NLTK (LOPER; BIRD, 2002). The Brazilian Portuguese stemming algorithm selected was RSLP (HUYCK; ORENGO, 2001). We implemented the lemmatization technique using Spacy (HONNIBAL; MONTANI, 2017), and TF-IDF is supported by the Scikit-learn library (PEDREGOSA et al., 2011). Additionally to the preprocessing techniques, we evaluated two supervised classifiers. We decided to focus on shallow classifiers due to the size of the educational corpus supporting the experiment. We chose the following algorithms:

– Random Forest: we selected the algorithm because it demonstrated robust performance in classifying English texts in different contexts (HARTMANN et al., 2019).

– SVM: we chose the algorithm because it is the most used classifier for text classification in Portuguese content, according to a systematic mapping study (SOUZA et al., 2018).

In order to evaluate the experiment, we implemented a k-fold cross-validation, described by BENGIO; GRANDVALET (BENGIO; GRANDVALET, 2005) as an intensive technique using all data available as training and testing examples. The technique implements K times the process of training a machine learning model, and 1/K fraction of the training examples is left out to test the model. We implemented a five-fold cross-validation, using one fold as the testing set and the four remaining folds for training the model. The five-fold is recommended by RODRIGUEZ; PEREZ; LOZANO (RODRIGUEZ; PEREZ; LOZANO, 2009) because it is less biased than a smaller number of folds. We opted to apply the stratified k-fold method available on Scikit-learn. This method ensures that each fold includes the same proportion of examples from each class included in the corpus (PRUSTY; PATNAIK; DASH, 2022).

In addition to k-fold cross-validation, we conducted multiple tests since the algorithms can produce slightly different results. We opted to run ten tests and showed the mean and the 95% confidence interval for the metrics obtained using repeated k-fold cross-validations. According to WONG; YEH (WONG; YEH, 2019), repeating k-fold cross-validation allows collecting more observations to obtain a reliable accuracy. However, the authors affirmed that "it is inappropriate to say the mean accuracy resulting from a larger number of replications will be more reliable".

Modern machine learning classifiers have parameters, named hyperparameters, that impact their performance and need to be set before training (WEERTS; MUELLER; VANSCHOREN, 2020). In this experiment, we implemented a script to evaluate the best hyperparameter setting for each combination of preprocessing techniques. Combining every aforementioned preprocessing technique, we evaluated the best setting for 12 different versions of preprocessed documents for both classifiers. We used the GridSearchCV method available in the Scikit-learn library to assess the hyperparameters. We used as a basis the hyperparameter sets evaluated by PROBST; BOULESTEIX; BISCHL (PROBST; BOULESTEIX; BISCHL, 2019) and WEERTS; MUELLER; VANSCHOREN (WEERTS; MUELLER; VANSCHOREN, 2020).

This experiment fits with multi-class classification because the corpus comprises four departments, and documents need to be classified in one of them. Therefore, we employed measures for multi-class classification indicated by SOKOLOVA; LAPALME (SOKOLOVA; LAPALME, 2009): Accuracy, Precision, Recall, and Fscore. We used macro-averaging to calculate the measures, treating all classes equally.

5.4 Experiment 3: Information Classification

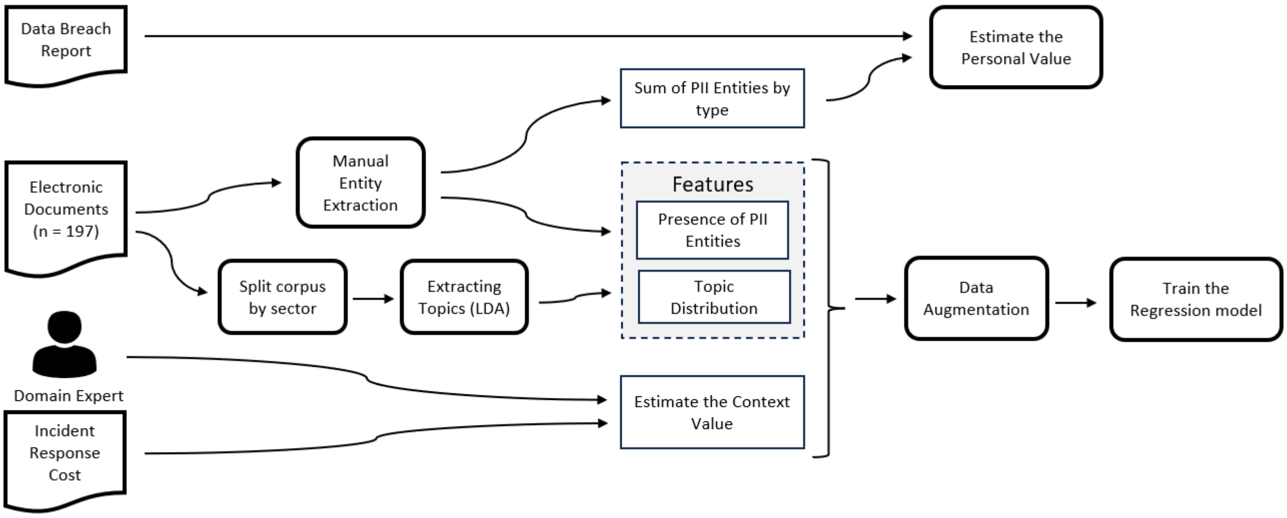
We designed an experiment to perform a complete evaluation of the proposed model supported by the previous experiments. The experiment's first stage, presented in Figure 11, comprises the activities to prepare the model for classifying the documents. In order to extract personal data features, we manually analyzed all documents in the corpus and identified the PII entities. The manual analysis produced two feature sets; one was used to estimate the personal data value, and the other was used to assess the context value. Additionally, using the training corpus, we generated one LDA model for each organization department and computed the topic distribution related to each document in the corpus. Ultimately, we implemented a data augmentation and trained a regression model. The activities were supported by internal information provided by domain experts and external reports.

5.4.1 Document value estimation

For training the machine learning regression model, we needed to value each document included in the corpus, and the proposed model estimates the document's worth based on two different values: (i) the value added by the presence of personal data; (ii) the value associated with the document content that we named as context value. As mentioned in Section 4.1, information valuation can be categorized into market-based, economic, and dimensional approaches (FLECKENSTEIN; OBAIDI; TRYFONA, 2023), and we used the market-based model to estimate the information value based on the cost associated with an information security incident.

For estimating the value associated with the presence of personal data in the document,

Figure 11 – The activities performed to prepare the model for value-based information classification.



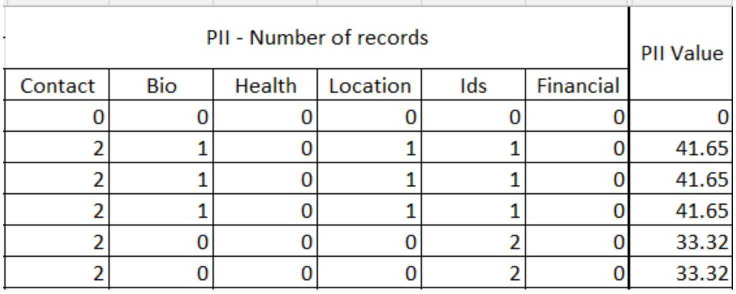
Source: Created by the author.

we used the IBM Cost of a Data Breach Report (IBM Security, 2023). The report presents data breach costs in several countries, including Brazil, and the business cost for each leaked record containing PII associated with a customer or an employee. Brazil has one of the lowest costs, representing around 27% of the global average compared to other countries. Although the report considers four process-related activities to estimate costs, in the experiment, we adjusted the cost of customer (USD 183.0) and employee (USD 181.0) PII based on the same ratio as the cost of data breaches in Brazil compared to the global average. Since using equivalent values to personal data and context values is essential, we considered the value of each customer and employee record to be USD 50.0, which corresponds to approximately 27% of the original customer and employee PII cost.

Based on the adjusted cost, we estimated the personal data value of each document by manually extracting the number of customer or employee PII records. It is important to reinforce that only entities associated with a person were registered in this activity; for example, we did not consider a phone or e-mail from the educational organization. The authors manually performed the entity extraction activity to create the training data by analyzing each document in the corpus. We considered the types of PII entities presented in Figure 7, which we evaluated in the first experiment. However, we believe it is incorrect to value the discovery of one entity in a document to USD 50.0. Thus, we divided the record value (USD 50.0) by the number of groups to determine the document value associated with personal data. In Figure 7, the 23 PII entities are arranged in six groups: contact, bio, health, location, finance, and ID. Next, we estimated the worth of each entity group at USD 8.33 and multiplied the entity with the highest frequency inside each group by this value.

Additionally, we disregard the entity type ’name’ for performing the personal value assessment since we believe its presence in a document does not mean it has worth. Figure 12 presents examples of the entity groups counting and the respective PII value calculated. The counters represent the most frequently discovered entity in each entity group. For example, in line two, the most frequent entity in the group "Contact" summed up two, and the most frequent entity in groups "Bio", "Location", and "IDs" appeared once; so the computed personal data value multiplied the number of each group occurrences by the worth assigned for the entity group record (USD 8.33). We are aware that different entities and groups could result in particular weight in value estimation. However, the scientific literature lacks a method to allow us to propose a more robust approach.

Figure 12 – Examples of personal data value estimation considering the most frequent entity for each group.



PII - Number of records						PII Value
Contact	Bio	Health	Location	Ids	Financial	
0	0	0	0	0	0	0
2	1	0	1	1	0	41.65
2	1	0	1	1	0	41.65
2	1	0	1	1	0	41.65
2	0	0	0	2	0	33.32
2	0	0	0	2	0	33.32

Source: Created by the author.

We determined the document context value using the numerical intensity approach described by Sajko et al. (SAJKO; RABUZIN; BAČA, 2006), which proposed the document categorization based on its importance to the organization. Based on the table presented by the authors, we adjusted the number of levels and the meaning of each intensity, considering the impact of information compromise (the original work considered the information loss). We defined four intensity levels, and we believe that for estimating the information value of documents, the higher the number of levels, the better. However, it was the first time that the organization participating in the experiment performed an analysis for projecting costs used to estimate information value, so we decided to use fewer levels. Table 8 presents the intensities and the modified descriptions.

As many documents have similar organizational functions, we clustered the 197 documents into 26 groups. In order to define the intensity of each document group, we presented our table's version to employees in leadership positions. We requested them to categorize the document groups they handled daily into one of the intensities. The categorization process involved the leader of each of the four departments that provided documents and a manager who knows the relevance of each group of documents for the organization, summing up five employees. Using this approach, two employees categorized the intensity of each group of documents.

After intensity categorization, we decided to measure interrater reliability before resolving divergences. Interrater reliability is used in scientific studies that demand humans to classify subjects or objects into predefined classes or categories and represents the extension to which humans' categorization coincides (GWET, 2014). We measured the interrater reliability by

Table 8 – The intensities and its descriptions presented to employee for document association.

Intensity	Meaning
1	The information has no economic value, so its compromise does not generate costs.
2	The information is interesting, and its compromise produces a small impact on the organization and/or short-term consequences. The compromise of information results in a small cost.
3	The information is important, and its compromise produces high costs and/or mid-term consequences. Information compromise has serious consequences for the organization.
4	The most valuable information whose compromise can have serious consequences for the organization. Information compromise has multiple impacts on the organization.

Source: Created by the author.

Table 9 – Number of groups and documents categorized into each intensity.

Intensity	Groups	Number of Documents
1	5	68
2	6	41
3	12	72
4	3	16

Source: Created by the author.

applying Cohen's kappa since it is a robust statistic test to evaluate the agreement level in studies counting on two raters (MCHUGH, 2012). The Cohen's kappa resulting was 0.16, indicating slight agreement according to Landis and Koch (LANDIS; KOCH, 1977). This result confirmed the subjectiveness problem related to human security classification presented in BERGSTRÖM; ÅHLFELDT (BERGSTRÖM; ÅHLFELDT, 2014) and ANDERSSON (ANDERSSON, 2023), and it might be expected in the current scenario since the organization does not implement an overall information classification policy. As it is necessary to associate each document group with a specific intensity to estimate its value, we scheduled specific meetings with the employees who categorized the documents to resolve divergences. Table 9 summarizes the number of groups and the total of documents categorized into each intensity.

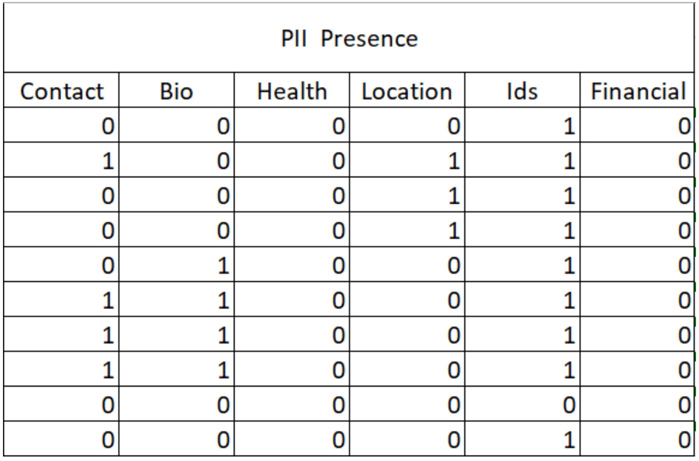
Although the document intensity offers a quantitative measure of each document, the proposed model demands a monetary value to use as the context value. In order to establish the relation between document intensity and its value, we deliberated on the information security sector to estimate the costs associated with information security incident response. We asked the managers to consider past information security incidents comprising information similar in intensity to those described in Table 8 and estimate the costs based on the involvement of the professionals in the incident response process and their corresponding remuneration. Although information security incidents produce multiple impacts in organizations (COUCE-VIEIRA; INSUA; KOSGODAGAN, 2020), we believe the costs associated with personnel offer an objective approach to estimating information value to evaluate the proposed model.

The approach resulted in ranges representing the hours demanded from each type of information security professional dealing with past information security incidents. We selected the central point of each range and multiplied it by the hourly cost of the professional. The estimated value is the sum of the costs of all professionals dealing with incidents related to different intensities. Since the value associated with personal data was estimated in dollars, we converted the context value associated with each document intensity from Brazilian Real to US Dollar (USD). In the end, we obtained three values associated with each document of the corpus: (i) the personal data value, (ii) the context value, and (iii) the document value, which results from the sum of the previous values.

5.4.2 Features extraction

The proposed model defines the extraction of two feature sets to represent the context information value and train the regression model. The first feature set corresponds to the PII entity's presence in each document, and we used the 23 PII entities manually extracted from documents to estimate the personal data value. However, we understand that the frequency of PII does not necessarily impact the context value, but its presence does. So, we did not consider the frequency of individual entities but rather the presence of one of the entities in each group. We also removed the entity 'Name' from the group 'Contact' for this activity. Thus, we extracted six features representing the presence of each of the six described groups. Figure 13 presents examples of the features extracted from documents representing the presence of distinct PII groups.

Figure 13 – Examples of features representing the PII presence in documents.



PII Presence					
Contact	Bio	Health	Location	Ids	Financial
0	0	0	0	1	0
1	0	0	1	1	0
0	0	0	1	1	0
0	0	0	1	1	0
0	1	0	0	1	0
1	1	0	0	1	0
1	1	0	0	1	0
1	1	0	0	1	0
0	0	0	0	0	0
0	0	0	0	1	0

Source: Created by the author.

The second feature set is based on topics extracted from the documents, and we decided to use the LDA model in this task because it is the most common topic modeling algorithm used to identify topics in text (ZHENG et al., 2023). For this activity, we preprocessed the documents before extracting the topics. Initially, we implemented a data cleansing script and removed noisy characters from conversion that were not manually removed. Next, we normalized the text to lowercase and lemmatize words. Finally, we removed the stopwords. NLTK and Spacy libraries supported the preprocessing script.

The next activity groups the preprocessed documents by each department. Considering the documents associated with each department, we implemented a stratified five-fold cross-validation approach considering the 26 document groups as classes. To ensure the success of this approach, we included at least five documents from each group in the corpus. We created LDA models for each department since previous tests extracting topics from all documents did not produce satisfactory results. The script was implemented using the Scikit-learn library.

A drawback of the LDA model is the need to specify the number of topics to be extracted as a model parameter. In the experiment, we implemented the approach proposed by ZHAO et al. (ZHAO et al., 2015) that defined a rate based on perplexity named RPC. This metric measures the capability of a model to generate the documents based on the topics extracted (ABDELRAZEK et al., 2023). We implemented an experiment configuration similar to ZHAO et al. (ZHAO et al., 2015) and built the LDA models using five-fold cross-validation comprising the documents of each department.

We evaluated different numbers of topics, starting with five and incrementing to five until it reached 70. We repeated the cross-validation test 30 times, assigning different random seeds for each repetition. It is important to highlight that the five-fold cross-validation implemented to calculate the RPC differs from the stratified K-fold approach implemented to create the LDA models. The results of the normalized RPC indicated 65 topics for the marketing model, 35 topics for the secretariat model, 40 topics for the diploma issuance model, and 45 topics for the document management model. We implemented the RPC approach using the Scikit-learn library.

After defining the number of topics, the next step was to generate the LDA models from the documents grouped by department. Thus, using the same library, we generated the LDA model for each set of documents, and for each model, we got the topic distribution of each document included in the set. The topic distribution represents the context features extracted from each document.

5.4.3 Training the regression model

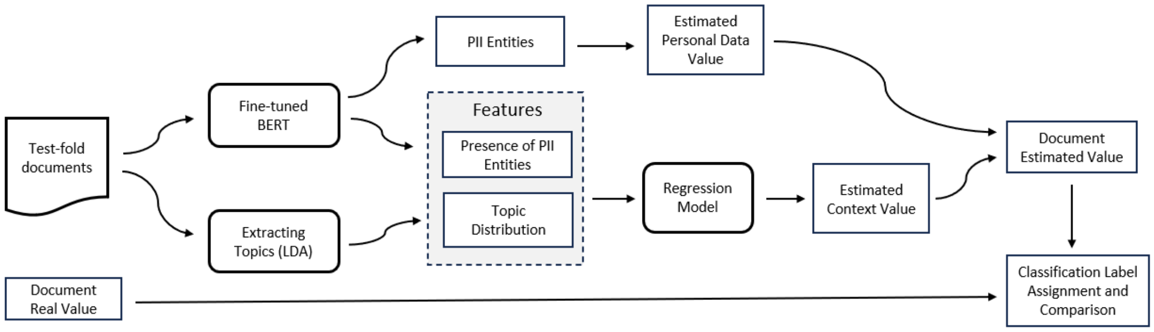
The last activity was to train a machine-learning model to learn the weights of the features extracted considering the document's personal data presence and topic distribution. However, we needed to deal with a corpus limitation before training the model. As shown in Table 9, there is a distinct number of documents related to each intensity. Since the intensity reflects the document value, which is used as the target class in the training, we needed to handle the unbalanced data problem. So, we decided to implement a data augmentation approach. Data augmentation encompasses strategies to increment new training examples without collecting new data (FENG et al., 2021b). One strategy for implementing data augmentation is over-sampling, which can randomly replicate minority class patterns or generate new patterns for minority classes (ELREEDY; ATIYA, 2019). We implemented a random over-sampling for the under-represented classes in the experiment using the Imbalanced-learn library (LEMAÃŽTRE; NOGUEIRA; ARIDAS, 2017).

The next step was choosing the regression models to be evaluated in the experiment. We based our decision on two criteria: (i) the performance of the regression model in small datasets in the metrics evaluated by FERNÁNDEZ-DELGADO et al. (FERNÁNDEZ-DELGADO et al., 2019); (ii) the availability of the algorithm in the Scikit-learn library since we used it for implementing this experiment's activity. Using these criteria, we selected Decision Tree (DT), Gaussian Process Regression (GPR), Gradient Boosting (GB), Random Forest (RF), and Support Vector Regression (SVR). Before training the regression models, we needed to define the hyperparameters for each classifier, so we used the method GridSearchCV available on Scikit-learn. We trained the regression model using the same stratified five-fold document sets resulting from the topic extraction activity. We also trained an individual model for each department using the five-fold approach, so GridSearchCV resulted in five indications of best parameters, and we used the most frequent for training each department's regression model.

5.4.4 Model Evaluation

After training the regression models, the subsequent stage evaluates the proposed model. Figure 14 presents the evaluation process's activities. The testing corpus is formed by the documents of the test fold implemented in model preparation, so all documents in the corpus are assessed in one fold.

Figure 14 – Process performed to evaluate the results produced for the model.



Source: Created by the author.

For discovering PII entities, we implemented the lowercase version of the fine-tuned BERT model trained in experiment 1, described in Section 5.2 as BERT$_{\text{uncased}}$. We selected the model because it achieved the best performance in the experiment. We also used the same RE approach implemented in experiment 1 to remove unrelated PII entities according to the distance window. Using the BERT model and the RE approach, we discovered PII entities in the documents included in the test fold. So, we implemented two Python scripts to use the entities for validating the proposed model. Considering the same entity groups shown in Figure 7 and not considering the entity 'name', the first script counted the frequency of each of the remaining entities and multiplied the most frequent entity in each group by the same value (USD 8.33) used in the preparation model. The sum of all multiplications results in the estimated personal data value. The second script records the presence of at least one entity in each group. As we proceeded in model preparation, the entity 'name' was removed from the 'Contact' group. The presence or absence of the entity groups comprises the first set of features used as input in the regression model.

After extracting features from PII entities, we applied the topic modeling task to extract the topic distribution from each document in the test fold. First, we implemented the same preprocessing techniques employed in the LDA model training. Next, we utilized the LDA model generated using the set of documents comprising the other four folds for each set of test-fold documents. The topic distribution and PII presence were used as the input in the regression model, which outputs the estimated context value for each document.

The personal data and context values were summed, resulting in the document's estimated value. At this point of the process, we estimated the value of each document using the proposed model. We also have the actual value of the document computed based on the approach described in Section 5.4.1. In order to evaluate security classification performance, we created information classification schemes for experimental scenarios associating ranges of values with sensitivity levels and defined a sensitivity label for each level. So, for each corpus document, we computed the corresponding label. The final outcome of the model is a label determining the information classification.

Thus, we employed metrics for evaluating multi-class classification to assess the performance of security classification. Based on SOKOLOVA; LAPALME (SOKOLOVA; LAPALME, 2009), we selected the following metrics: Accuracy, Precision, Recall, and Fscore. The metrics computation used the weighted average parameter available in the Scikit-learn library to handle class imbalance. Additionally, as we implemented regression models, we calculated the Root Mean Square Error (RMSE), considering actual and estimated values. We also computed the RMSE considering the real and estimated PII values.

A hindrance to assessing multi-class classification metrics is the lack of an information classification scheme defining values for each sensitivity label to use as a reference. Thus, as mentioned earlier, we decided to establish experimental scenarios for distinct information classification schemes. Initially, we analyzed several information classification policies that are publicly available, and we ascertained that organizations usually implement information classification schemes comprising three, four, or five levels. We also analyzed the RFC describing information security policies from three companies, which all fit in one of these classification schemes (NICOLLS, 2002).

After defining the number of levels to be applied in the evaluation, we arranged two scenario sets. Each scenario set consisted of three classification schemes, involving one scheme for each number of levels considered in the evaluation. So, we evaluated the proposed model using six independent scenarios, comprehending the three information classification schemes. We defined the labels' description based on the most frequent in the information classification policies analyzed.

In order to define the ranges of values for each scenario, we considered two criteria: (i) the public label must be related to a very low value aiming to avoid the assignment of a public label to documents containing distinct PII entities; (ii) for the most part, value ranges must incorporate the majority of documents. In the second criterion, we considered that the approach explained in Section 5.4.1 estimated a context value equal to or lower than USD 1,100 for most documents. Table 10 presents the range of values considered in the experiment to classify documents and analyze the model performance. For example, in the first experimental scenario related to the three-level information classification scheme, if the document's worth is estimated to be USD 10.0 or lower, the model assigns the label "Public". If the estimated value for the document is superior to USD 10.0 and equal to or lower than USD 300.0, the model assigns the label "Internal". Finally, the document receives the label "Confidential" if its estimated value exceeds USD 300.0.

Table 10 – The six experimental scenarios defined to assign the sensitivity label to a document. We evaluated two scenarios for each classification scheme considering distinct ranges of values (values in US dollars).

Levels	Label	Experimental scenario set #1		Experimental scenario set #2	
		Lower Bound	Upper Bound	Lower Bound	Upper Bound
3	Public	-	10.0	-	15.0
	Internal	10.01	300.0	15.01	450.0
	Confidential	300.01	-	450.01	-
4	Public	-	10.0	-	15.00
	Internal	10.01	100.0	15.01	300.0
	Confidential	100.01	400.0	300.01	800.0
	Highly Confidential	400.01	-	800.01	-
5	Public		10.0		15.0
	Internal	10.01	100.0	15.01	300.0
	Confidential	100.01	400.0	300.01	800.0
	Secret	400.01	1,000.0	800.01	1,500.0
	Top Secret	1,000.01	-	1,500.01	-

Source: Created by the author.

6 RESULTS AND DISCUSSIONS

This chapter presents the results achieved in the three experiments designed. Section 6.1 describes the performance of LSTM and BERT models to discover PII entities in documents from organizations related to educational and health sectors. Section 6.2 presents the performance achieved by the TF-IDF approach to classifying documents into four departments of the educational organization. Finally, Section 6.3 outlines the proposed model performance to estimate the value and assign the sensitivity label to documents.

6.1 Experiment 1: PII Entities Discovery

This section evaluates the information extraction performance in discovering PII entities following the experiment design described in Section 5.2. Aiming for a better organization, we divided the explanation into four sub-sections. In Section 6.1.1, we describe the performance of the four trained models considering only the application of NERC without validating the entities and implementing RE. Section 6.1.2 presents the final result of the RE approach implemented for identifying only PII entities. It is important to reinforce that the NERC result considered all entities found in the testing corpus related to the set of 23 personal data illustrated in Figure 7, and after performing the RE approach, we considered only the entities found in the testing corpus that should be protected. Section 6.1.3 analyzes the impact of NERC and RE for discovering documents containing PII entities. In this evaluation, we focus on identifying documents that include personal data. Finally, in Section 6.1.4, we discuss the overall results and present the study's limitations.

6.1.1 NERC Models Performance

After performing the NERC approaches, we generated the first result set. Table 11 presents the number of entities correctly, incorrectly, and partially correct recognized for each model. The table also shows how many entities were missed by each model and the total of false positives (spurious). The total number of entities in the 80 documents comprising the testing corpus is 906.

The results show that all models missed a significant quantity of entities in the documents. The model that best performed in discovering entities was $\text{BERT}_{\text{uncased}}$. Although the number of missed entities is quite worrying when applying the models for discovering entities for information security proposals, analyzing the entities identified in each document makes it possible to verify that most missed entities were related to addresses data found in the documents unrelated to PII. Additionally, we identified that the models did not recognize several people's names included in longer documents.

Table 11 also presents that BERT and LSTM models identified only a few spurious entities, and it is very relevant in this study's context since the objective of the NERC is to identify documents that need to be protected because they hold personal data. The small number of spurious entities supposes that the models generate few false positives. This is relevant because it reduces the number of documents that need to be analyzed by the information security personnel.

Table 11 – Table presents the number of correct, incorrect, partially correct, missed, and spurious entities for each of the four evaluation types. This table considers all entities identified in the documents, even those unrelated to personal data.

Model	Evaluation Type	Correct	Incorrect	Partial	Missed	Spurious
$\text{BERT}_{\text{cased}}$	Strict	486	68	0	352	12
	Exact	507	47	0	352	12
	Partial	507	1	46	352	12
	Type	532	22	0	352	12
$\text{BERT}_{\text{uncased}}$	Strict	573	77	0	256	37
	Exact	585	65	0	256	37
	Partial	585	0	65	256	37
	Type	637	13	0	256	37
$\text{LSTM}_{\text{cased}}$	Strict	565	24	0	317	46
	Exact	572	17	0	317	46
	Partial	572	0	17	317	46
	Type	582	7	0	317	46
$\text{LSTM}_{\text{uncased}}$	Strict	545	45	0	318	20
	Exact	550	40	0	318	20
	Partial	550	0	40	318	20
	Type	588	2	0	318	20

Source: Created by the author.

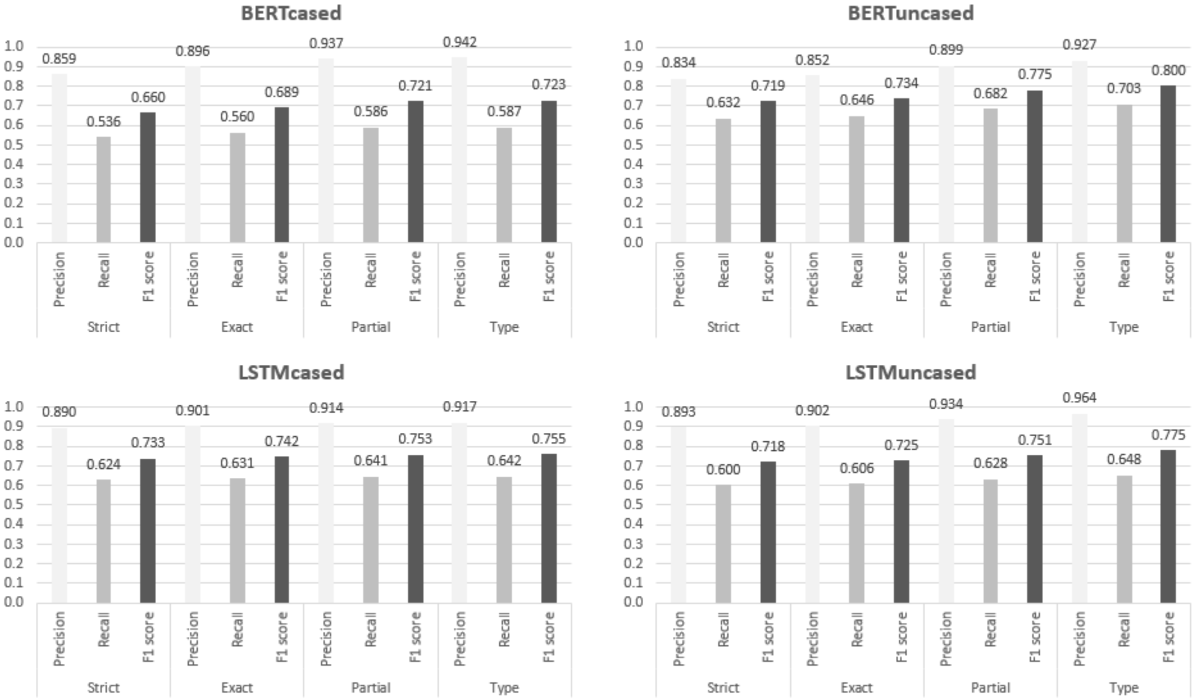
Figure 15 and Table 12 present the selected metrics considering the documents related to the health and education sectors. Considering the document sample assessed in this study and the F1-score metric, $\text{BERT}_{\text{uncased}}$ in the Type approach was the model that achieved the best performance. The F1 score was 0.8. Two models scored 0.775, the second higher, $\text{BERT}_{\text{uncased}}$ and $\text{LSTM}_{\text{uncased}}$. The difference was that $\text{BERT}_{\text{uncased}}$ achieved the metric in the Partial approach and $\text{LSTM}_{\text{uncased}}$ in the Type approach.

Table 12 – Precision, Recall, and F1-score results for each model considering the set of 80 documents from the health and education sectors. The best F1 scores are highlighted in bold.

Named Entity Recognition and Classification (NERC) - Health and Education documents												
Model	Strict			Exact			Partial			Type		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
$\text{BERT}_{\text{cased}}$	0.859	0.536	0.660	0.896	0.560	0.689	0.937	0.586	0.721	0.942	0.587	0.723
$\text{BERT}_{\text{uncased}}$	0.834	0.632	0.719	0.852	0.646	0.734	0.899	0.682	**0.775**	0.927	0.703	**0.800**
$\text{LSTM}_{\text{cased}}$	0.890	0.624	0.733	0.901	0.631	0.742	0.914	0.641	0.753	0.917	0.642	0.755
$\text{LSTM}_{\text{uncased}}$	0.893	0.600	0.718	0.902	0.606	0.725	0.934	0.628	0.751	0.964	0.648	**0.775**

Source: Created by the author.

Figure 15 – Precision, Recall, and F1-score results for each model considering the set of 80 documents from the health and education sectors.



Source: Created by the author.

6.1.2 Relation Extraction for PII Discovery

The proposed process applied the NERC outcome to the next stage comprising validation functions and RE rules. Unlike the NERC approach that aimed to recognize any of the 23 entity types assessed in this study, even if the entity is unrelated to a person, this second stage should identify only PII entities. This stage's result should indicate if the information security personnel must protect a specific document. Table 13 presents the first result of this stage. The table shows the number of correctly, incorrectly, partially correct, missed, and spurious entities identified for each evaluated model. Since this stage must recognize only PII-related entities, the total number of entities in the 80 documents included in the testing corpus is 566.

We observed that the number of spurious entities increased for all models because an entity recognized as correct in the NERC assessment can be identified as spurious in the RE evaluation if unrelated to personal data. For example, if NERC recognized a company's address and the RE rules did not exclude it, it was counted as spurious since there is no association with a person. The analysis of Table 13 makes it possible to verify that the implemented rules failed in removing some entities recognized in the NERC stage. The lowest number of spurious entities was identified by BERT$_{\text{cased}}$, summing up 22, followed by BERT$_{\text{uncased}}$, which summed up 48. Therefore, in our assessment, the BERT-fine-tuned models detected the fewest spurious entities.

On the other hand, the number of missed entities decreased. As we mentioned in Section 6.1.1, the NERC approach missed many entities unrelated to personal data, which were not considered in this evaluation. Regarding missed entities, BERT-fine-tuned models also performed better than LSTM. However, unlike the previous result, the BERT$_{\text{uncased}}$ summed up 89, the lowest number of missed entities. The BERT$_{\text{cased}}$ model missed 121 PII entities.

Table 13 – Table presents the number of correct, incorrect, partially correct, missed, and spurious entities for each of the four evaluation types after performing entity relation. This table considers only entities related to personal data.

Model	Evaluation Type	Correct	Incorrect	Partial	Missed	Spurious
BERT$_{\text{cased}}$	Strict	389	56	0	121	22
	Exact	407	38	0	121	22
	Partial	407	0	38	121	22
	Type	423	22	0	121	22
BERT$_{\text{uncased}}$	Strict	426	51	0	89	48
	Exact	439	38	0	89	48
	Partial	439	0	38	89	48
	Type	464	12	0	89	48
LSTM$_{\text{cased}}$	Strict	366	12	0	188	103
	Exact	374	4	0	188	103
	Partial	374	0	4	188	103
	Type	370	8	0	188	103
LSTM$_{\text{uncased}}$	Strict	350	13	0	203	53
	Exact	354	9	0	203	53
	Partial	354	0	9	203	53
	Type	359	4	0	203	53

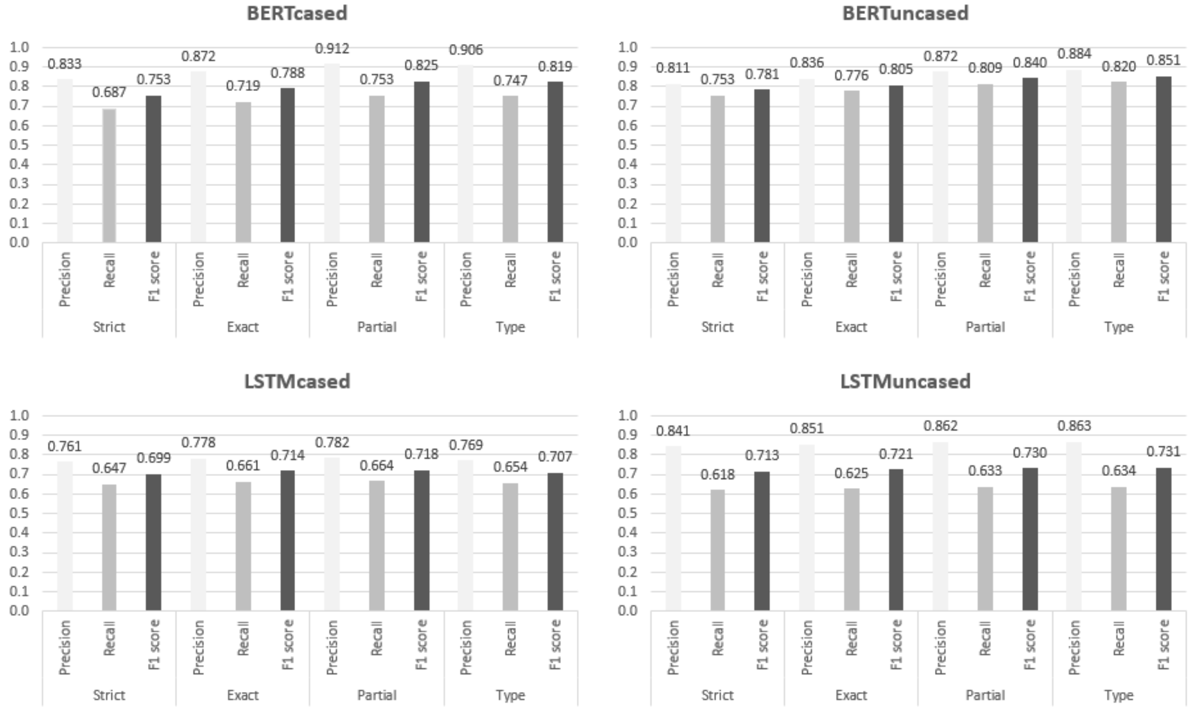
Source: Created by the author.

Figure 16 and Table 14 present the metrics obtained after performing RE. We observed that BERT models achieved a superior F1 score considering only PII entities. On the other hand, LSTM models worsen the F1-score metric. Again, the BERT$_{\text{uncased}}$ model in the Type approach achieved the best performance. It is worth noting that both Precision and Recall achieved metrics pretty close, both superior to 0.8, which suggests that the number of missed and spurious entities was low. The F1-score calculated for this approach was 0.851. BERT$_{\text{uncased}}$ model also achieved an F1-score of 0.84 in the Partial approach, which was the second highest. Following BERT$_{\text{uncased}}$, the BERT$_{\text{cased}}$ was the model to achieve the highest F1-score. The Partial approach reached an F1 score of 0.825. However, in this case, the model did not present a balance between Precision and Recall since it achieved a Precision of 0.912, the best for all models, but the Recall was 0.753.

As mentioned in Section 5.1.1, two document sets were evaluated in the experiment. Tables 15 and 16 present the metrics for each set of documents individually. As in the consolidated result, the BERT$_{\text{uncased}}$ model achieved the best performance in each set individually. The model obtained an F1-score of 0.904 for the educational documents. The Partial approach achieved this result and was slightly higher than the 0.901 obtained by the Type approach. Based on this number, we can assume that BERT$_{\text{uncased}}$ assigned incorrect labels for a small set of recognized entities. Despite that, we think the model's performance was pretty good for educational documents, considering the set of distinct PII entities in the evaluation.

The BERT$_{\text{uncased}}$ model obtained an F1-score of 0.828, considering only documents related

Figure 16 – Precision, Recall, and F1-score results for each model after performing relation extraction considering the set of 80 documents from the health and education sectors.



Source: Created by the author.

to the health sector. The result also demonstrates that the BERT$_{\text{uncased}}$ model achieved the best result in identifying PII entities in education documents than health sectors. As we mentioned in Section 5.2.2, the health organization provided us with digitized documents based on physical forms, and we observed that BERT models did not perform well in this type of document. We observed that BERT$_{\text{cased}}$ also obtained superior performances in education documents. This performance was not noticed in LSTM models since the cased version obtained a better F1 score in health documents.

6.1.3 Undiscovered Documents

An issue of traditional discovery approaches, mainly based on regular expressions and dictionaries, is the significant number of discovered entities associated with false positives. On the other hand, we must avoid missing documents containing PII. So, considering our sample of 80 documents, we analyzed the results of each model evaluated to determine if they might miss documents containing PII entities. Table 17 shows the number of documents in which the NERC models recognized no entity and how much it represents from the sample evaluated. The table also presents the same data after performing the RE approach.

The table demonstrates that the BERT$_{\text{uncased}}$ model did not miss any document in the NERC approach. In the cased model of BERT, the NERC approach resulted in the loss of one document. The LSTM resulted in missing three documents in both models, cased and uncased.

Table 14 – Precision, Recall, and F1-score for each model after performing relation extraction rules considering the 80 documents from the health and education sectors. The best F1 scores are highlighted in bold.

Results after performing relation extraction in Health and Education documents												
Model	Strict			Exact			Partial			Type		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
BERT$_{\text{cased}}$	0.833	0.687	0.753	0.872	0.719	0.788	0.912	0.753	**0.825**	0.906	0.747	0.819
BERT$_{\text{uncased}}$	0.811	0.753	0.781	0.836	0.776	0.805	0.872	0.809	**0.840**	0.884	0.820	**0.851**
LSTM$_{\text{cased}}$	0.761	0.647	0.699	0.778	0.661	0.714	0.782	0.664	0.718	0.769	0.654	0.707
LSTM$_{\text{uncased}}$	0.841	0.618	0.713	0.851	0.625	0.721	0.862	0.633	0.730	0.863	0.634	0.731

Source: Created by the author.

Table 15 – Precision, Recall, and F1-score for each model after performing relation extraction rules considering 40 documents from the education sector. The best F1 scores are highlighted in bold.

Results after performing relation extraction in Education documents												
Model	Strict			Exact			Partial			Type		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
BERT$_{\text{cased}}$	0.842	0.827	0.835	0.885	0.869	0.877	0.906	0.890	0.898	0.885	0.869	0.877
BERT$_{\text{uncased}}$	0.833	0.863	0.848	0.862	0.893	0.877	0.888	0.920	**0.904**	0.885	0.917	**0.901**
LSTM$_{\text{cased}}$	0.589	0.726	0.651	0.614	0.756	0.677	0.614	0.756	0.677	0.589	0.726	0.651
LSTM$_{\text{uncased}}$	0.744	0.726	0.735	0.768	0.750	0.759	0.774	0.756	0.765	0.756	0.738	0.747

Source: Created by the author.

After applying entity relation rules, we obtained a more significant number of undiscovered documents. BERT$_{\text{uncased}}$ presented the best scenario, and the number of missed documents increased from zero to three, representing 3.75% of the sample. There are two reasons for the increment in the number of missed documents after applying RE in the entities recognized by BERT$_{\text{uncased}}$.

The first reason pertains to the NERC approach's recognition of entities unrelated to PII. For example, one missed document comprised four entities: three were related to personal data, and one was associated with the city of the health organization, which had no relation to personal data. When we performed the NERC approach in this document, the entity related to the city was recognized, so the model discovered only one entity. However, when RE was applied, as the name of the city was found alone in the document, the entity was excluded, resulting in no entities remaining.

The second reason is associated with the design of RE rules. The rules considered the distance between entities when deciding whether to keep or remove them. In two cases, the BERT$_{\text{uncased}}$ model missed PII entities, resulting in recognizing entities pretty far from each other, and the consequence was the exclusion of both. One example was a document comprising nine entities for NERC evaluation, two related to personal data positioned at the beginning of the document and seven associated with address and contact entities of the educational organization. The BERT uncased model recognized two entities, the person's name (one of the two personal data) and the phone number of the educational organization, at the end of the document. Due to

Table 16 – Precision, Recall, and F1-score for each model after performing relation extraction rules considering 40 documents from the health sector. The best F1 scores are highlighted in bold.

Results after performing relation extraction in Health documents												
Model	Strict			Exact			Partial			Type		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
BERT$_{\text{cased}}$	0.828	0.628	0.714	0.864	0.656	0.746	0.916	0.695	0.790	0.917	0.696	0.791
BERT$_{\text{uncased}}$	0.801	0.706	0.750	0.823	0.726	0.772	0.865	0.763	0.810	0.883	0.779	**0.828**
LSTM$_{\text{cased}}$	0.891	0.613	0.726	0.901	0.621	0.735	0.909	0.626	0.741	0.905	0.623	0.738
LSTM$_{\text{uncased}}$	0.905	0.573	0.702	0.905	0.573	0.702	0.919	0.582	0.712	0.933	0.590	0.723

Source: Created by the author.

Table 17 – The number of documents containing PII that each model would not discover, considering the entities identified by the NERC models and after performing the RE approach.

Model	NERC		Relation Extraction	
	Total Docs	% of Total	Total Docs	% of Total
BERT$_{\text{cased}}$	1	1.25	4	5
BERT$_{\text{uncased}}$	0	0	3	3.75
LSTM$_{\text{cased}}$	3	3.75	9	11.25
LSTM$_{\text{uncased}}$	3	3.75	11	13.75

Source: Created by the author.

the distance between the recognized entities, both were excluded by entity relation rules. These two reasons also impacted the other models.

6.1.4 Discussion and Limitations

In this section, we added discussions about the results achieved in the experiment, associating them with the discovery of personal data, the experiment's primary motivation. We also explore the results in the context of recognizing fine-grained entities and the application of a generic training corpus in different domains. Finally, we present some limitations related to the experiment designed.

In the NERC evaluation, the BERT model fine-tuned with lowercase documents achieved the best performance with an F1-score of 0.8. Although BERT obtained the best performance in the NERC evaluation, the LSTM$_{\text{uncased}}$ model also obtained an F1-score of 0.775. The sample of documents evaluated showed that BERT and LSTM performed better for uncased models. Despite this study considering only 24 entities, a smaller number compared to other fine-grained studies, we understand that BERT and LSTM should be considered for recognizing fine-grained entities.

In a broader analysis of the NERC results, we observed that the Type approach obtained the higher F1-score for all models evaluated, and this indicates that models correctly recognize the entities' tags but fail to identify the entities' boundaries. This result is positive, considering using NERC to identify PII entities in documents. From the information security perspective, knowing the correct type of PII for most entities is very relevant, even if some were not framed accurately. The primary reason is that some data types are more valuable than others. For example, many government and bank sites utilize the federal ID number to identify the user, so cyber criminals constantly target this type of data, which should receive special attention from information security teams.

A central point in the experiment was evaluating the RE contribution to discard entities unrelated to personal data but identified in the NERC stage. So, the token distance approach was applied to obtain some context of the text to decide whether an entity refers to PII. Again, $\text{BERT}_{\text{uncased}}$ associated with the Type metric achieved the best performance. Considering only PII-related entities, $\text{BERT}_{\text{uncased}}$ achieved an F1-score of 0.851, improving the metric to 0.05 if compared to NERC evaluation. The better performance in the F1 score resulted from a significant improvement in the Recall since many entities missed in the NERC stage were related to non-PII entities.

On the other hand, we verified that the precision metric achieved a lower result after RE because many non-PII entities were not removed using the token distance approach. Based on this, we understand that improving the precision metric through fine-tuning RE rules is possible. We also understand this is only possible by increasing the number of documents in the testing corpus because crafting rules based on the established corpus could generate a bias.

The training corpus used in the experiment comprised documents from different sectors, and we were careful not to include documents from the domains related to the testing corpus. Cross-domain NER is a well-known challenge due to the difficulty for a trained model to generalize entity recognition for distinct domains (JIA; LIANG; ZHANG, 2019; LIU et al., 2021). The results demonstrated that when fine-tuned with a generic-labeled corpus, $\text{BERT}_{\text{uncased}}$ recognized entities in distinct domain documents. In Section 6.1.2, we presented the results for each domain individually, and $\text{BERT}_{\text{uncased}}$ performed an F1-score above 0.8 in both domains, despite a difference of 0.07 between them.

Although the $\text{BERT}_{\text{cased}}$ and LSTM models did not perform best, we believe the approaches should be evaluated with a training corpus comprising a more significant number of entities. Conversely, as mentioned in Section 5.2.3, the CRF-trained model did not obtain a positive performance in an initial assessment, and we decided to remove it from the experiment. Based on the initial evaluation, we understand that CRF models do not perform well in recognizing fine-grained entities in multiple domains.

Discovering personal data entities is crucial for organizations to overcome the lack of knowledge related to PII stored in their infrastructure. It is also needed because organizations can be penalized and fined for violating laws and regulations. Based on the testing corpus consisting of 80 documents, results demonstrated that only the $\text{BERT}_{\text{uncased}}$ model recognized entities in all documents in the NERC evaluation. After applying RE, all models failed to identify PII entities; consequently, organizations would not protect some documents. On the other hand, results demonstrated the identification of a small number of spurious entities, so a real-world application should not misidentify a large number of documents. Considering the number of documents stored in organizations, a small number of false positives is very important to ensure supervision from security teams and that an organization does not waste resources and time protecting unnecessary documents.

To conclude, it is essential to highlight that the thesis proposed model relies on PII recognition to estimate personal data and context values. Therefore, this experiment is essential to evaluate the viability of the model. Based on the results achieved in this experiment, we understand that the NERC and RE approaches evaluated in this experiment contribute positively to the model's success since they allow the recognition of most entities. Although the NERC and RE missed a few entities, the approach implemented in the model experiment minimizes the impact since it considers the presence and frequency of an entity group.

Although we based the experiment on a well-known methodology to obtain consistent results, it is essential to present some limitations related to this study. The first limitation is the number of industry sectors providing documents to establish the testing corpus. The experiments evaluated the NERC approaches in documents provided by two industry sectors. As every industry sector usually uses proper language and handles different document structures, this context limits our evaluation.

The number of entities evaluated in the experiment is also a limitation. We annotated 24 entities for training the model and evaluated the model considering 20 entities included in document sets provided by organizations. In real-world applications, documents might contain entities not comprised in the experiment, and models should deal with them. Finally, the number of annotations related to some entities is also a limitation. The documents in the training corpus are unbalanced, and some entities were more frequently observed than others. This limitation can primarily impact ID-related entities because some have similar features, and the NERC approaches could create a bias.

6.2 Experiment 2: Document Classification

As we mentioned in Section 4.3, initially, we performed an attempt to extract only one topic distribution considering all corpus documents. The experiment did not perform satisfactorily, and we proposed extracting a distinct topic distribution for each department to improve the proposed model's results. This enhancement demands the classification of the documents into departments, and the proposed model presents a text classification activity before topic modeling. So, we designed an experiment to evaluate text classification performance using the traditional BoW approach and analyze if the enhancement is viable. We assessed two machine learning classifiers and multiple combinations of preprocessing techniques.

Tables 18 and 19 present the results achieved with Random Forest and SVM classifiers, respectively. For each metric, we present the average computed in the ten repetitions and the lower and upper limits of the confidence interval between brackets. In the column "Preprocessing", we present the combination of preprocessing techniques ordered according to their implementation. For example, in the line presenting the description "Lw + Stop + Lemma," we lowercased the documents, removed the stopwords, and finally converted the remaining terms to their lemmas. In the line showing "No Preprocessing", no preprocessing techniques were implemented.

Table 18 – Results for the mean and its 95% confidence interval for each preprocessing combination employing the Random Forest model.

Preprocessing	Random Forest			
	Accuracy	Precision	Recall	F1
No Preprocessing	0.975 (0.966,0.983)	0.977 (0.969,0.986)	0.975 (0.968,0.983)	0.976 (0.968,0.984)
Lw	0.978 (0.971,0.985)	0.979 (0.972,0.986)	0.979 (0.972,0.985)	0.979 (0.972,0.985)
Stop	0.979 (0.973,0.986)	0.981 (0.975,0.987)	0.980 (0.974,0.985)	0.980 (0.975,0.986)
Stem	0.969 (0.960,0.978)	0.973 (0.965,0.982)	0.971 (0.962,0.979)	0.972 (0.963,0.980)
Lemma	0.979 (0.972,0.987)	0.981 (0.973,0.989)	0.980 (0.973,0.987)	0.980 (0.972,0.988)
Lw + Stop	0.980 (0.974,0.986)	0.981 (0.974,0.987)	0.981 (0.975,0.986)	0.980 (0.974,0.986)
Lw + Stem	0.975 (0.969,0.981)	0.977 (0.971,0.983)	0.976 (0.971,0.981)	0.976 (0.971,0.982)
Lw + Lemma	0.972 (0.965,0.980)	0.975 (0.967,0.983)	0.973 (0.967,0.980)	0.974 (0.967,0.981)
Lw + Stop + Stem	0.978 (0.973,0.984)	0.981 (0.976,0.986)	0.978 (0.973,0.983)	0.979 (0.974,0.984)
Lw + Stop + Lemma	0.977 (0.969,0.986)	0.980 (0.971,0.988)	0.977 (0.969,0.986)	0.978 (0.970,0.987)
Stop + Stem	0.975 (0.966,0.983)	0.978 (0.970,0.986)	0.975 (0.968,0.983)	0.976 (0.969,0.984)
Stop + Lemma	0.980 (0.973,0.987)	0.983 (0.977,0.990)	0.980 (0.974,0.986)	0.981 (0.975,0.988)

Preprocessing techniques abbreviations: Lowercase conversion (Lw); Stop word removal (Stop); Stemming (Stem); Lemmatization (Lemma).

Source: Created by the author.

Analyzing Table 18 and using the accuracy metric, we can affirm that the confidence intervals revealed that no preprocessing combination has a significant advantage over the others since there is an overlapping in the confidence interval for all combinations. The interpretation of the accuracy metrics achieved by the SVM classifier shown in Table 19 allows us to assert that the absence of preprocessing techniques was outperformed by two combinations, with confidence intervals not overlapping at the 95% level. The confidence interval upper limit when no preprocessing was implemented was 0.983. In the evaluation that we only lowercased documents, the confidence interval lower limit computed was 0.987, and when we lowercased, removed stopwords, and lemmatized documents in a row, the lower limit was 0.986. Considering the 95% confidence interval, the combination that outperformed the absence of preprocessing also surpassed the implementation of the solo stop word removal technique, which computed an upper limit of 0.984.

The TF-IDF approach significantly performed in classifying the corpus used in the experiment. Some corpus characteristics can influence the performance achieved. The corpus comprises four distinct departments of the educational organization, and some documents can contain quite different terms. For example, documents provided by the marketing department included commercial language, while documents from other departments used formal writing.

Furthermore, different documents associated with a specific department can hold similar structures and use the same terms. For example, among the content provided by Document Management are distinct types of contracts, and the Diploma Issuance department provided several documents incorporating a partial course's curriculum.

Table 19 – Results for the mean and its 95% confidence interval for each preprocessing combination applying the SVM model.

Preprocessing	SVM			
	Accuracy	Precision	Recall	F1
No Preprocessing	0.984 (0.983,0.985)	0.988 (0.987,0.989)	0.983 (0.982,0.984)	0.985 (0.984,0.986)
Lw	0.989 (0.987,0.990)	0.990 (0.989,0.992)	0.988 (0.987,0.990)	0.989 (0.988,0.991)
Stop	0.982 (0.980,0.984)	0.984 (0.982,0.986)	0.982 (0.980,0.983)	0.983 (0.981,0.985)
Stem	0.983 (0.979,0.988)	0.986 (0.981,0.990)	0.985 (0.980,0.989)	0.985 (0.980,0.989)
Lemma	0.986 (0.984,0.988)	0.989 (0.987,0.991)	0.985 (0.983,0.987)	0.986 (0.984,0.989)
Lw + Stop	0.986 (0.984,0.989)	0.986 (0.984,0.988)	0.986 (0.983,0.988)	0.985 (0.983,0.988)
Lw + Stem	0.986 (0.983,0.989)	0.988 (0.985,0.991)	0.988 (0.985,0.990)	0.988 (0.985,0.990)
Lw + Lemma	0.986 (0.983,0.990)	0.987 (0.984,0.990)	0.987 (0.984,0.989)	0.987 (0.984,0.990)
Lw + Stop + Stem	0.982 (0.978,0.986)	0.983 (0.979,0.987)	0.982 (0.979,0.986)	0.983 (0.979,0.986)
Lw + Stop + Lemma	0.988 (0.986,0.991)	0.989 (0.987,0.990)	0.988 (0.986,0.991)	0.988 (0.986,0.990)
Stop + Stem	0.980 (0.974,0.986)	0.982 (0.977,0.988)	0.981 (0.975,0.988)	0.982 (0.976,0.988)
Stop + Lemma	0.986 (0.984,0.988)	0.989 (0.986,0.991)	0.985 (0.983,0.987)	0.987 (0.984,0.989)

Preprocessing techniques abbreviations: Lowercase conversion (Lw); Stop word removal (Stop); Stemming (Stem); Lemmatization (Lemma).

Source: Created by the author.

Based on the results achieved in this experiment, we consider it viable to implement the TF-IDF approach to classifying documents by department for the educational corpus. The text classification task will forward most documents to the correct organizational department so the proper topic modeling model will extract the topic distribution.

6.3 Experiment 3: Information Classification Model

This section discusses the results achieved for the proposed model. We analyzed two implementations of the proposed model, considering the corpus provided by the educational organization and the four departments involved in the experiment. Aiming for a better organization, we divided the results achieved in each implementation into different sections. Section 6.3.1 shows the results of extracting a single topic set from all corpus documents. Section 6.3.2 considers the extraction of one distinct topic set for each department. The document classification activity supports this implementation. Finally, Section 6.3.3 discusses the results and contributions of the proposed model and presents some limitations.

6.3.1 The Proposed Model using a Single Topic Set

In the first moment, we considered the whole corpus to extract the topic modeling distribution and train the regression models. This approach generated LDA models comprising documents from the four organizational departments and the five-fold cross-validation. We applied the RPC method, and the number of topics configured in LDA was 70. After generating the LDA models, we extracted each document's topic distribution in the training corpus and trained the regression models. Table 20 presents the results achieved for the regressors. We bolded the best accuracy in each scenario considering the three information classification schemes evaluated. RF obtained the best accuracy in four scenarios, and DT performed better in two.

Table 20 – The results achieved by the proposed model extracting a single set of topics from the 197 documents. The best accuracy achieved for each scenario is highlighted in bold.

Regressor	Levels	Scenario set #1				Scenario set #2			
		Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
Decision Tree	Three	0.680	0.774	0.680	0.675	0.711	0.761	0.711	0.711
	Four	0.680	0.774	0.680	.0.675	**0.711**	0.761	0.711	0.711
	Five	0.680	0.774	0.680	0.675	**0.706**	0.761	0.706	0.705
Gaussian	Three	0.360	0.643	0.360	0.280	0.355	0.643	0.355	0.270
	Four	0.355	0.643	0.355	0.270	0.355	0.643	0.355	0.270
	Five	0.350	0.643	0.350	0.260	0.355	0.562	0.355	0.270
Gradient Boosting	Three	0.609	0,708	0.609	0.585	0.609	0.737	0.609	0.594
	Four	0.497	0.723	0.497	0.545	0.503	0.793	0.503	0.567
	Five	0.310	0.803	0.310	0.440	0.467	0.743	0.467	0.525
Random Forest	Three	**0.736**	0.780	0.736	0.730	**0.731**	0.774	0.731	0.729
	Four	**0.731**	0.793	0.731	0.736	0.706	0.795	0.706	0.729
	Five	**0.701**	0.810	0.701	0.731	0.701	0.791	0.701	0.723
SVR	Three	0.513	0.639	0.513	0.553	0.538	0.648	0.538	0.567
	Four	0.472	0.648	0.472	0.540	0.503	0.682	0.503	0.548
	Five	0.401	0.717	0.401	0.488	0.482	0.652	0.482	0.510

Source: Created by the author.

Table 21 presents the RMSE computed for each regression model. The table shows that the GB model achieved the lower error (582.52) among all regressors. This contrasts with the results in Table 20, which does not associate the model with the best accuracy in any scenario. This divergence occurs because the GB model computed a slight error for more documents while RF and DT calculated a superior error for fewer documents. RF achieved the second-lowest error (689.29) and the best accuracy in four scenarios. On the other hand, the Gaussian regressor obtained the worst RMSE (1,121.50) and the worst result in five scenarios. Interestingly, the GB model obtained the worst result in the only scenario in which Gaussian did not achieve the worst accuracy.

6.3.2 The Proposed Model using a Topic Set by Department

Table 22 presents the results after extracting a distinct topic set from each department. The table shows that DT achieved the best accuracy for all experimental scenarios created for the

Table 21 – The RMSE computed for each regressor considering a single topic set.

Regressor	RMSE
Decision Tree	719.97
Gaussian	1,121.50
Gradient Boosting	**582.52**
Random Forest	689.29
SVR	897.87

Source: Created by the author.

Table 22 – The results achieved by the proposed model extracting a distinct topic set for each organizational department. The best accuracy achieved for each scenario is highlighted in bold.

Regressor	Levels	Scenario set #1				Scenario set #2			
		Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
	Three	**0.843**	0.864	0.843	0.848	**0.838**	0.861	0.838	0.843
Decision Tree	Four	**0.817**	0.858	0.817	0.834	**0.838**	0.864	0.838	0.844
	Five	**0.817**	0.858	0.817	0.834	**0.807**	0.831	0.807	0.813
	Three	0.662	0.704	0.662	0.677	0.652	0.728	0.652	0.671
Gaussian	Four	0.571	0.716	0.571	0.632	0.465	0.746	0.465	0.510
	Five	0.359	0.710	0.359	0.417	0.444	0.639	0.444	0.470
	Three	0.822	0.841	0.822	0.826	0.797	0.815	0.797	0.802
Gradient Boosting	Four	0.802	0.859	0.802	0.826	0.756	0.828	0.756	0.788
	Five	0.721	0.876	0.721	0.790	0.736	0.805	0.736	0.765
	Three	0.832	0.857	0.832	0.832	0.802	0.839	0.802	0.807
Random Forest	Four	0.777	0.852	0.777	0.809	0.731	0.852	0.731	0.779
	Five	0.690	0.869	0.690	0.767	0.701	0.819	0.701	0.743
	Three	0.359	0.207	0.359	0.262	0.350	0.200	0.350	0.250
SVR	Four	0.359	0.207	0.359	0.262	0.348	0.199	0.348	0.253
	Five	0.359	0.207	0.359	0.262	0.348	0.199	0.348	0.253

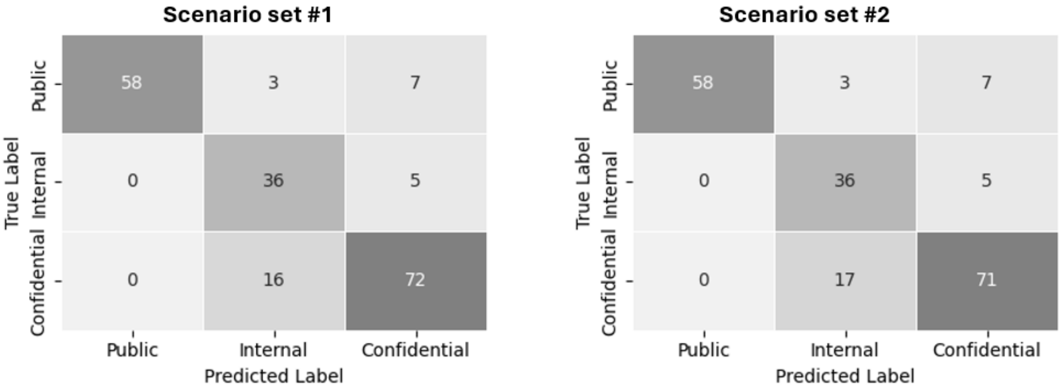
Source: Created by the author.

evaluation. The DT classifier achieved an accuracy exceeding 0.8 across all scenarios, while the GB method matched this performance in three scenarios, and RF did so in only two. The results allow us to affirm that the DT model achieved the best performance in this experiment.

We also analyzed the confusion matrices produced by regression models to verify the differences in the security classification, considering the six information classification schemes designed for the experiment. As Gaussian and SVR did not perform satisfactorily, we excluded them in the following analysis and focused on the three regression classifiers that achieved the best performances. The DT model's confusion matrix comprising the three-level scheme is presented in Figure 17. Figure 18 shows the confusion matrix for the GB model, and Figure 19 for the RF model.

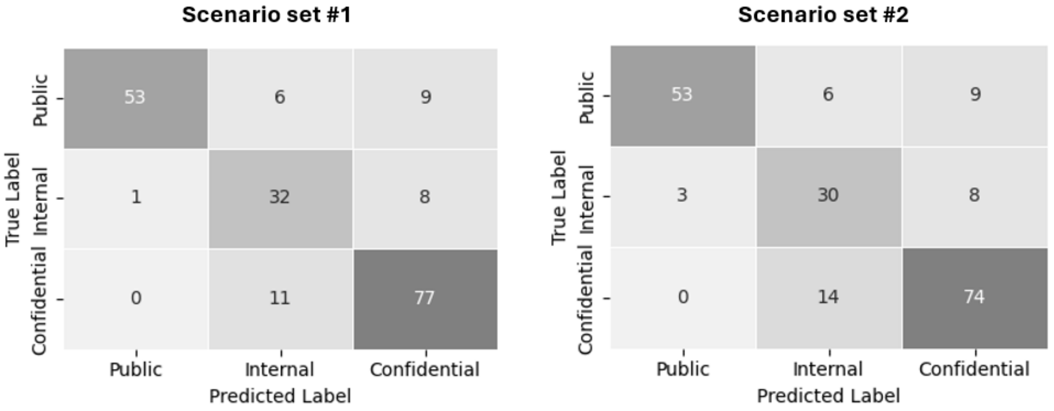
In order to support the confusion matrices' analysis, it is essential to highlight that an under-classified document creates problems regarding confidentiality and integrity properties, and information classified at a higher level produces an issue related to the availability property. The discussions regarding confusion matrices focused on confidentiality and availability properties, considering each information classification scheme evaluated. We decided to restrict the discussion to these information security properties because the integrity discussion could produce repetitive analysis.

Figure 17 – The Decision Tree confusion matrices for the scenario sets #1 and #2 considering the three-level information classification schemes.



Source: Created by the author.

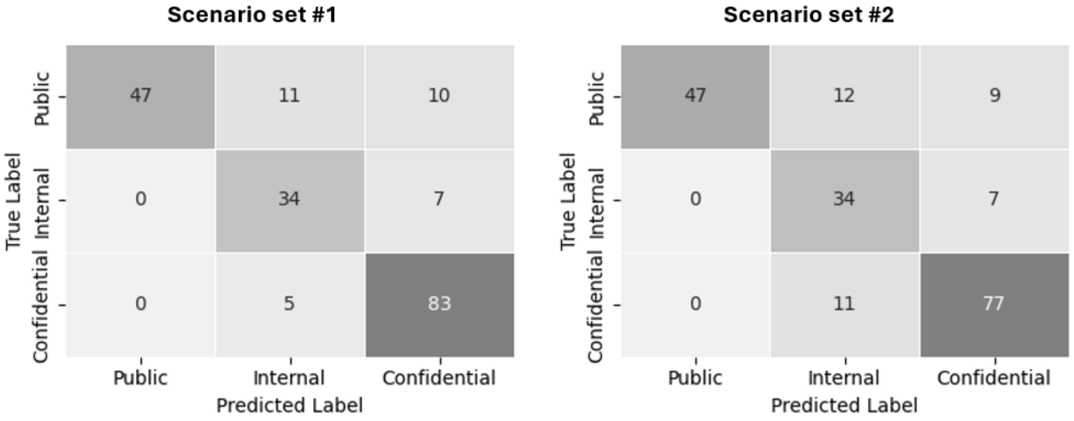
Figure 18 – The Gradient Boosting confusion matrices for the scenario sets #1 and #2 considering the three-level information classification schemes.



Source: Created by the author.

Examining the three-level scheme matrices, we can affirm that RF was the model with the lowest number of documents under-classified. So, RF was the model that least impacted confidentiality, although it did not achieve the best accuracy in either scenario. A specific concern regarding confidentiality in the information classification process is the public compromise of an organization's document. If a document is classified as "Public", employees are not restricted from sending it out of the organization's limits. Analyzing the matrices, we verified that GB assigned the label 'Public' for one internal document in scenario #1 and for three internal documents in scenario #2. Considering the concern regarding misclassifying internal and confidential documents as 'Public' in the three-level schemes, the confusion

Figure 19 – The Random Forest confusion matrices for the scenario sets #1 and #2 considering the three-level information classification schemes.



Source: Created by the author.

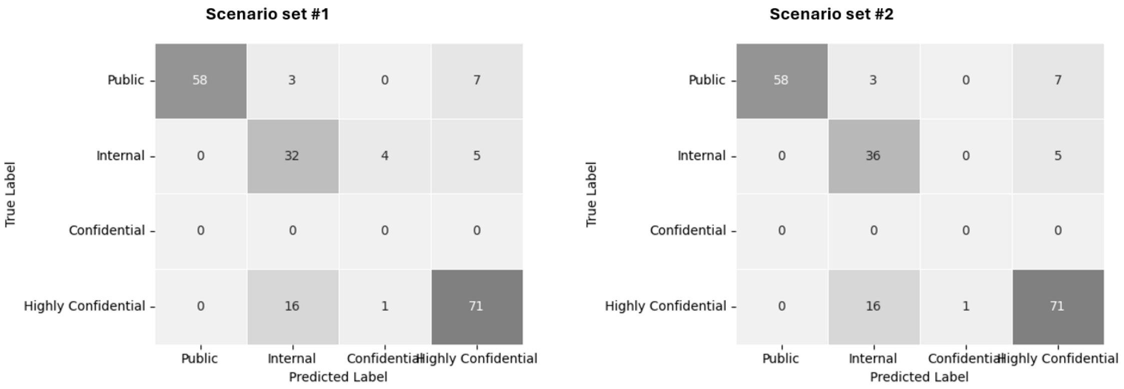
matrices indicate that DT and RF models are the best alternatives for implementing information classification in the experimental scenarios, considering confidentiality property.

Analyzing the three-level matrices allows us to affirm that GB and RF overestimated the value of more documents compared to the number of documents undervalued. The same did not occur with DT; the model undervalued 16 and 17 documents in scenarios #1 and #2, and it overestimated 15 documents in each scenario. Unlike confidentiality property, there is no difference if a model wrongly classified a document one or more levels higher because, in both cases, the document is inaccessible to authorized people. RF was the model causing the most significant impact on availability because authorized employees could not access 28 documents in scenario #1 and 26 documents in scenario #2. DT was the model producing the least impact on information availability.

Figure 20 presents the confusion matrix for the four-level information classification scheme for the DT model, and Figure 21 shows the confusion matrix associated with the GB model. Figure 22 illustrates the confusion matrix related to the RF model. The interpretation of matrices regarding the confidentiality property identified that RF under-classified the lowest number of documents in scenario #1, summing up ten documents. However, it performed worse in scenario #2 and under-classified 25 documents, while DT assigned a lower classification for 17 documents, the best number among the classifiers. As we have not changed the value ranges for the first level of both four-level information classification schemes, GB continued classifying documents that should be protected as "Public".

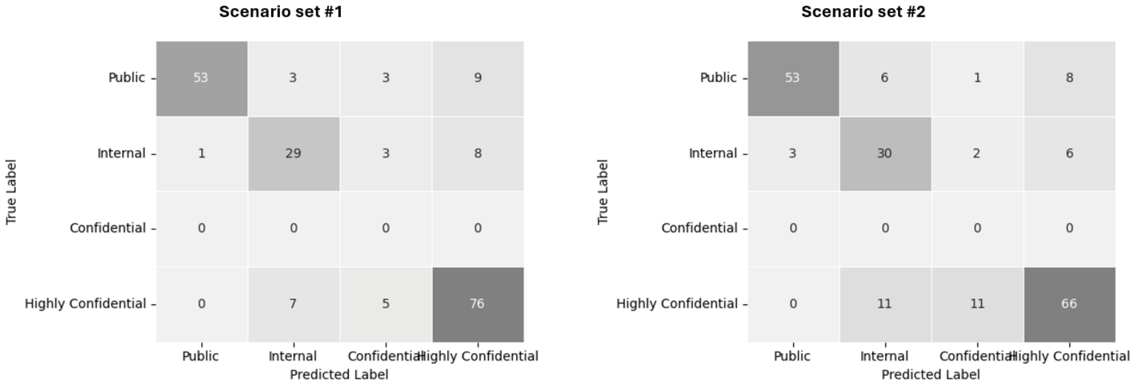
By examining the availability aspect, DT also overestimated the value of fewer documents in the four-level scheme scenarios than GB and RF. DT wrongly classified 19 and 15 documents at a higher level in scenarios #1 and #2, respectively. In contrast, GB and RF overestimated the value of 26 and 34 documents in scenario #1 and 23 and 28 documents in scenario #2. In summary, DT achieved the best result in classifying the corpus documents in both scenarios, considering the availability property, and performed best in one scenario when the objective was to ensure confidentiality. The RF obtained the best result in classifying documents regarding confidentiality in the other scenario.

Figure 20 – The Decision Tree confusion matrices for the scenario sets #1 and #2 considering the four-level information classification schemes.



Source: Created by the author.

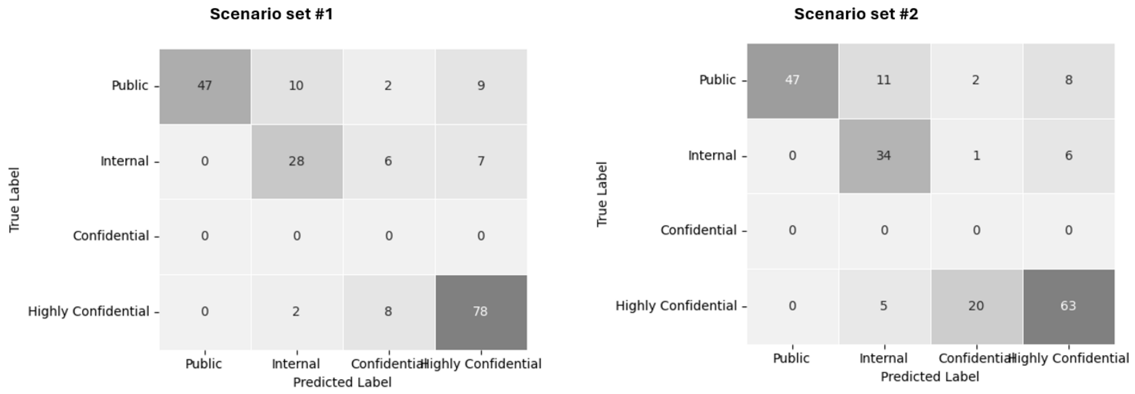
Figure 21 – The Gradient Boosting confusion matrices for the scenario sets #1 and #2 considering the four-level information classification schemes.



Source: Created by the author.

By evaluating the five-level confusion matrices, the finer-grained scheme, it is possible to identify meaningful issues in the experiment results. Although the classification performance of the DT model has not presented a significant change in the five-level scheme scenarios, the model classified 16 "Top Secret" documents as "Internal" in scenario #1, the lowest level excepting the "Public". This classification would allow all employees to access highly restricted information. In scenario #2, this issue is mitigated since the superior limit of the classification levels "Internal", "Confidential", and "Secret" is higher than in scenario #1, and only 16 documents are considered "Top Secret". In scenario #2, GB and RF correctly classified 15 documents as "Top Secret" while DT classified 13, and all models classified one "Top Secret" document as "Internal".

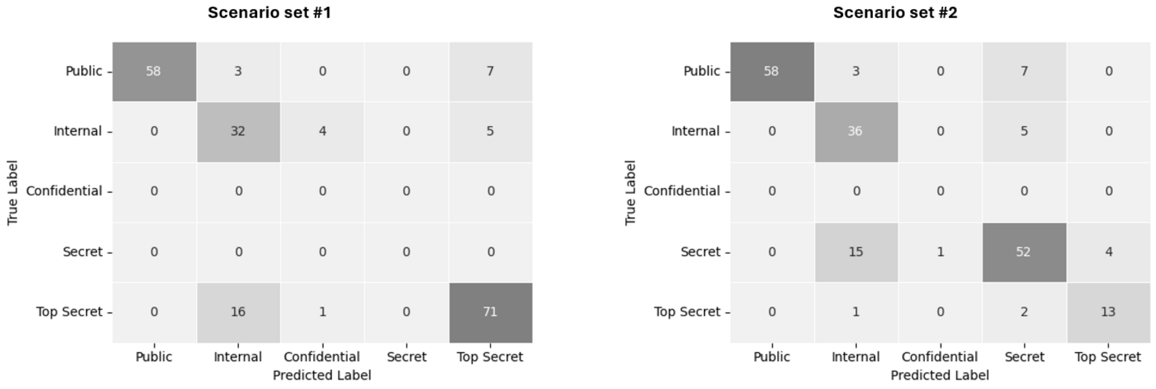
Figure 22 – The Random Forest confusion matrices for the scenario sets #1 and #2 considering the four-level information classification schemes.



Source: Created by the author.

The impact of document value estimation in the availability property did not change in scenario #1 of the five-level scheme compared to the four-level scheme. In the first scenario of the five-level scheme, we added a sensitivity level associated with documents with a value between US$ 400.00 and US$ 1,000.00, and no document had its value estimated in this range. Including a new level in scenario #2 impacted models' performances, making more documents inaccessible to authorized employees. In the DT, inaccessible documents increased from 15 to 19. For GB and RF, the number of inaccessible documents reached 25 and 34, respectively, an increase of 2 and 6 documents. The confusion matrix presenting the DT result related to the five-level information classification scheme is shown in Figure 23. The confusion matrix for the GB model is presented in Figure 24, and for the RF model in Figure 25.

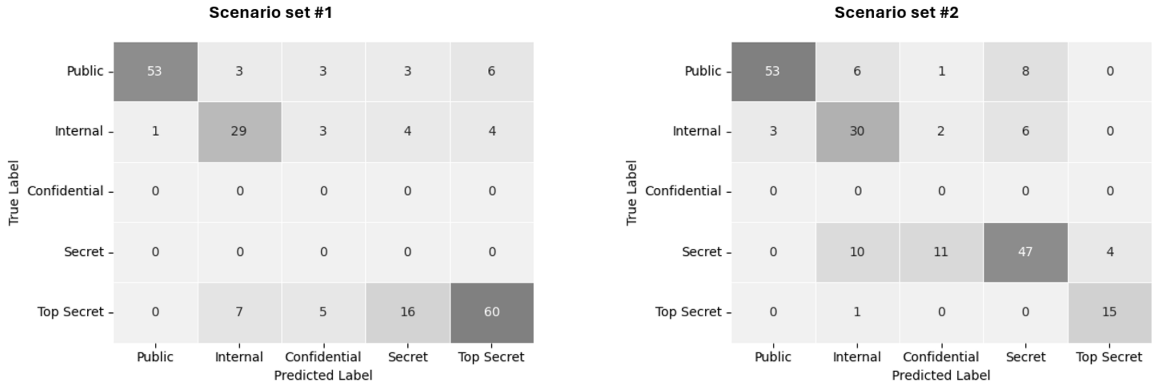
Figure 23 – The Decision Tree confusion matrices for the scenario sets #1 and #2 considering the five-level information classification schemes.



Source: Created by the author.

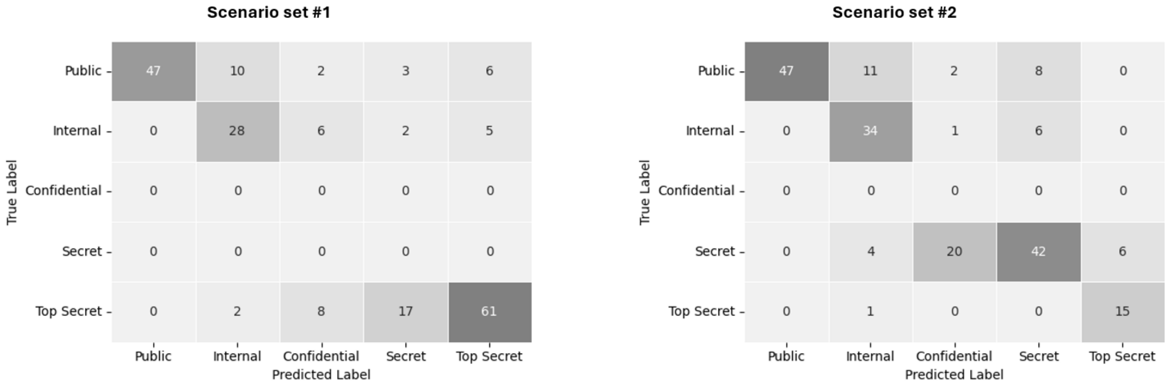
Implementing regression models allows us to demonstrate the RMSE computed in the experiment for the three models discussed. Random Forest (RF) obtained the lower RMSE, computing an error of 509.03. The RMSE computed for Decision Tree (DT) and Gradient

Figure 24 – The Gradient Boosting confusion matrices for the scenario sets #1 and #2 considering the five-level information classification schemes.



Source: Created by the author.

Figure 25 – The Random Forest confusion matrices for the scenario sets #1 and #2 considering the five-level information classification schemes.



Source: Created by the author.

Boosting (GB) were 596.45 and 548.42 respectively. So, despite RF's computation of the lowest error among regressors, it computed worse metrics than DT in the experiment. On the other hand, we verified the impact of a higher RMSE in DT in more fine-grained schemes; the most visible case was verified in the five-level scheme where many "Top Secret" documents were classified as "Internal" in scenario #1.

The trained model to estimate the context value considered the presence of PII in the document, and we can verify the accuracy in identifying the six distinct groups. Table 23 presents the accuracy obtained by the NERC model to discover PII considering the corpus documents. Most PII groups obtained an accuracy higher than 0.8; the only exception was the Health group, formed by only one personal data, which achieved 0.68. It is important to note that the educational corpus used in the experiment did not contain any document with personal data related to the health group, so the information extraction approach recognized false positives. Although we could evaluate the model's performance by performing feature engineering to remove the feature related to the presence of Health group entities, we think the

Table 23 – The accuracy of the NERC approach in identifying PII entities associated with each group.

PII Group	Accuracy
Contact	0.84
Bio	0.81
Health	0.68
Location	0.92
IDs	0.93
Financial	0.94

Source: Created by the author.

proposed model must generalize, so we decided to keep the feature in the experiment. We also computed the RMSE associated with the personal data value, which registered 16.85. Comparing the estimated PII values to the manually extracted value, we identified that the NERC model recognized many false positives. Hence, the approach based on the maximum PII in each group overvalued the personal data value in many documents.

6.3.3 Discussion and Limitation

Despite the wrongly classified documents, we consider the proposed model to have achieved positive results and offered many advantages over the traditional human-based approach. As described in different research (BERGSTRÖM; ANTERYD; ÅHLFELDT, 2018; AKSENTIJEVIĆ; TIJAN; AGATIĆ, 2011; ANDERSSON, 2023), and reinforced in this study, employees can disagree about the sensitivity of documents due to the subjectiveness of the task. As a result of the lack of consensus, employees can assign different classifications to similar documents. This study demonstrated that the agreement level between the organization's employees in leadership positions measured by Cohen's kappa was 0.16. On the other hand, by implementing the DT regressor, the proposed model achieved an accuracy higher than 0.80 in all experimental scenarios. Based on these two metrics, it is possible to infer that human classification would result in more errors in the information classification than the automated approach in the organization where this study was performed.

The results achieved by the DT classifier also demonstrated consistency considering different information classification schemes. By implementing the proposed model, an organization can change its classification scheme by only modifying the ranges of values associated with each classification label. Although classification schemes are usually stable, organizations can change them. Modifying information classification schemes can produce several challenges for an organization using a human-based approach, such as retraining employees and reclassifying all documents.

Based on the analysis of the matrices, it is possible to verify that a trained regression model can impact the information security properties differently. Thus, it is recommended that an organization executes an experiment to evaluate the performance of distinct regressors. When reviewing the experiment results, if no classifier emerges as the top performer in accurately estimating information value across diverse security properties, the organization must establish which property takes precedence to identify the optimal regression model.

It is pertinent to highlight that even supervised learning approaches designed to classify information based on sensitivity classes also demand massive effort if the classification scheme changes. Supervised approaches primarily depend on corpora comprised of a massive number of documents for training (OTHMAN; FAIZ; SMAÏLI, 2022), and human labeling is time-consuming (COLACE et al., 2014). Thus, if an organization previously implemented a supervised learning approach based on sensitivity classes and decides to change the information classification scheme, it must re-label all documents in the training corpus impacted by the scheme modifications. On the other hand, the proposed model presents a straightforward approach for organizations to modify their classification scheme with minimal effort. If the value-based approach was implemented, the organization only needs to adjust the value ranges associated with each sensitivity label in the information classification scheme. This scenario evidenced a specific advantage of the value-based approach.

Another specific advantage of the proposed model is the consideration of the presence of personal data in the information classification process. As demonstrated in the study of Loré et al. (LORÈ et al., 2023), many public documents contain personal data, and it configures a data breach. Information classification models that primarily rely on text classification to assign the sensitivity class to a document do not assign the proper weight to PII entities. This approach is employed in most information classification proposals, and some examples are presented in HUANG; CHIANG; CHANG (HUANG; CHIANG; CHANG, 2018), ALZHRANI et al. (ALZHRANI et al., 2019), and HEINTZ et al. (HEINTZ et al., 2022). Conversely, the proposed model implemented information extraction to discover PII entities, so it applies a dedicated focus on dealing with personal data. The approach implemented in the proposed model can be considered an advantage in comparison with text classification-based proposals since it considers the relevance of personal data in the document value. Moreover, clustering the entities proved positive because the information extraction implementation achieved significant results.

The RMSE computation is another advantage of the proposed model because an organization can calculate the metric using the same document set in the model preparation stage. The organization can use the RMSE computed to define the value ranges associated with each sensitivity class. Using the metric to establish the value ranges, an organization can adjust the information classification scheme to improve performance metrics.

As discussed in Section 3.3, one challenge within the information classification process is document reclassification. However, existing proposals in related work have not addressed how to effectively handle this challenge, which significantly impacts manual and automated information classification. The human-based approach demands that every time a document suffers a modification, the information owner receives a notification and reevaluates if the classification level needs adjustment. The automated approach makes the process easier since a system can reclassify the document after every change. Nevertheless, adopting supervised learning for information classification may present certain obstacles, such as re-annotating and retraining the classifier to adjust the model to the new criteria associated with the sensitivity levels. The constant retraining of a classifier can produce difficulties that make real-world organization implementation unpractical.

While the proposed model does not directly address the reclassification challenge, its design makes it easier to reclassify documents. The primary reason is the transparency regarding the document features used for estimating the document value. The model implementation can store the features related to the value assessment of each document, ensuring the traceability of the information classification process. For example, considering the personal data aspect, it is possible to verify if the presence or frequency of the entity groups changed after each document modification. If the frequency changes, the personal data value must be updated. In case the presence of PII suffered modification, the context value needs to be reassessed.

In another situation requiring reclassification, if documents from a department containing specific terms have their value reassessed, it is possible to retrain the department's regression model and reevaluate documents associated with the same topics. This last example highlights how extracting topics from each department's documents bolsters the model's effectiveness in addressing the reclassification challenge. Lastly, it is imperative to underscore that while the proposed model does not directly focus on document reclassification, its design distinguishes it from other supervised approaches by effectively handling this challenge during implementation. The benefits of transparency extend beyond reclassification, as it facilitates the implementation of an auditable information classification system. This system allows for understanding which document features were considered when assigning specific sensitivity labels.

The proposed model exhibits algorithm-agnostic characteristics, which is advantageous as it allows for adaptable implementation, capable of evolving alongside employing new regression and NLP algorithms. Throughout our experiments, we assessed multiple algorithms across various tasks and discussed their respective performances. In particular, the model's versatility extends beyond the algorithms evaluated, providing flexibility for future implementations.

It is not possible to compare the metrics obtained in the experiment with other studies due to a lack of baselines. The absence of baselines is justified since it is rare to have a corpus containing confidential documents available publicly. As we are using documents written in Brazilian Portuguese, it is even more difficult. Although we can not compare the experiment results to related work since they used different corpora, we verified that some published studies achieved an accuracy lower than 0.8 (ENGELSTAD et al., 2015b; ZARDARI; JUNG, 2016). Others obtained results between 0.8 and 0.9 (LIANG et al., 2019; HAMMER et al., 2015). So, even though we can not compare the results based on the metrics published in other studies, we understand that the results achieved with the proposed model are positive. This study also differs from others in the information classification research topic because it evaluates a corpus formed by Brazilian Portuguese documents. Based on this, the metrics obtained using a Brazilian Portuguese corpus are similar to those obtained in some studies evaluating English corpora.

Although we have taken care to articulate clear justifications for every decision made in our experiments, we consider it necessary to discuss some limitations. The proposed model was evaluated using documents from a single organization associated with the educational sector. We believe assessing the model using distinct corpora comprising multiple industry sectors is important. However, working with real-world documents is usually challenging since scholars depend on organizations to provide information. This limitation did not allow for measuring the model performance based on multiple corpora, and consequently, it does not make generalization possible.

The corpus used to base the experiment is also a limitation since it is unbalanced and comprises only 197 documents, which is considered small. Based on the set of documents received from the organization, we could only increment the corpus with documents comprising similar structures, which could produce a bias. So, despite the size of the corpus used in the study, which can impact the results, we think assessing a model using an unbiased corpus containing real-world documents is valuable.

Another limitation is the imbalance of documents' intensities. As presented in Table 9, employees estimated high values for only a few documents. Regardless, we implemented a script to balance documents before training the regression classifier; the over-sampling method did not produce variability in topic distribution, which might impact the document value estimation. We were aware of the unbalanced corpus, which reflects the real world since organizations store more low-value documents than high-value ones. In the '90s, Peltier (PELTIER, 1998) confirmed this statement, asserting that nearly 90 percent of an organization's information needs to be accessed by employees or is publicly available. Thus, we understand that creating an experiment similar to the real-world scenario is valuable to measure the model performance.

The number of intensity levels used to estimate documents' context value is also a limitation. Since we defined only four intensities, we assigned the same context value for many documents. As a result, no documents were associated with some sensitivity levels of the information classification schemes used to evaluate the experiment. So, an expanded number of intensity levels would result in more significant variability in documents' context values, which would be positive for evaluating more granular information classification schemes. Finally, we designed the criteria for the experimental scenarios. In a real-world implementation, an organization can assign the document values, and based on the computed RMSE and the information classification scheme, it can define the values related to each sensitivity level. However, before evaluating the model, we defined the value ranges for all information classification schemes in the experiment.

7 CONCLUSION

This thesis proposed an information classification model to assign documents' sensitivity based on the information value. The study evaluated the application of NLP to extract features related to personal data and content context and used them to estimate the value of information. Scholars have assessed intangible assets for various purposes, such as valuing intellectual property and estimating risks. However, to the best of our knowledge, this approach has not been explored by research in the information classification field. We implemented three experiments to evaluate the proposed model. The first experiment assessed approaches to recognizing PII entities in Brazilian Portuguese documents, and the second experiment analyzed the use of text classification to classify documents in distinct departments. These two experiments supported the third one, which evaluated the proposed model.

The first experiment demonstrated that NERC models achieve significant performance in discovering personal data and should be considered an alternative to traditional techniques. The NERC models evaluated in the study recognized only a few spurious entities, meaning the approaches produced a few false positives. It is very positive from the information security perspective since the excess of false positives could create an overload of work to protect the documents, much of it unnecessarily. In opposition, we also demonstrated that the trained models missed entities in the NERC and RE stages. Although RE reduced the number of missed entities, we presented that a consequence of missing some entities is that documents containing personal data would not be discovered.

The experiment evaluating text classification indicated that the TF-IDF correctly classifies organization documents by department. We tested a unique LDA model comprising topics from documents of multiple departments, but this implementation achieved poor performance. Thus, text classification is necessary to create a specific LDA model for each department, and the experiment indicated that the inclusion of the NLP task in the information classification process does not significantly impact the result.

The final experiment presented that the value-based approach is a possible alternative to traditional approaches employed in information classification. Our evaluation is based on six experimental scenarios designed by the authors. Using the scenarios, we evaluated five regression models. The study demonstrated that the DT performed better in most experimental scenarios, and its classification metrics presented significant consistency, so we can affirm that DT achieved the best performance in classifying documents based on the corpus used in this study.

The incorporation of PII entities for estimating information worth was also positive. The experiment results demonstrated that the information extraction approach achieved a significant performance in identifying the presence of PII entities. If the information classification scheme defines a low superior limit for the public level, discovering PII entities automatically restricts document access. Based on the results achieved, we can answer the research question and confirm that the experiment demonstrated that applying information extraction and topic modeling to extract PII entities and context features is a relevant approach to estimating information value and classifying documents. Moreover, it is possible to relate internal and external costs to existing documents, aiming to estimate the value of new ones by applying NLP tasks and regression methods.

This study also demonstrated that automating information classification can be an attractive alternative for organizations still implementing the process. During the experiment, we measured the interrater reliability of the organization's employees in classifying documents according to their sensitivity, and the result indicated a slight agreement. A possible consequence could be the assignment of inconsistent sensitivity labels in a human-based information classification process. On the other hand, the proposed model ensures that NLP tasks assign the same sensitivity label to similar content.

This thesis endeavors to provide a dual contribution. In the scientific field, it has contributed to the information security and information classification research fields by publishing two academic papers. The first paper applied the SLR method to present a broad view of the current scenario involving the application of text-mining tasks in the cybersecurity area (IGNACZAK et al., 2021). We also published a second paper describing the performance of information extraction in discovering Brazilian-related PII (IGNACZAK et al., 2023). Together, these publications demonstrated the originality and depth of this research. We also produced a third paper presenting the information classification model performance, currently under review in a scientific journal. The thesis also provides a technological contribution because the startup founded by the author is implementing a software solution grounded in the proposed model. The software, named Data Hunter®, is a data discovery machine learning-base solution that recognizes Brazilian PII entities. Moreover, the startup intends to develop an information classification solution based on the model proposed in the thesis.

We also identified opportunities for future work while carrying out this study. We propose expanding the number of documents and including organizations related to other industry sectors to assess the information classification model and the PII discovery approach. Although information value is a long-term research topic, we verified the need for studies assessing the worth of personal data. Due to recent regulations and laws, research comprising personal data has gained attention in different areas, and the value of different types of PII is relevant to support experiments. Future work can also evaluate the implementation of topic-filtering approaches to reduce the number of features to train the regression models. It is important because the method to define the number of topics for the LDA model can estimate a higher number of topics for corpora comprising a higher diversity of documents.

The application of RE to differentiate PII entities from others unrelated to personal data and avoid false positives in the discovery is also a relevant topic that needs further study, and we consider it relevant to evaluate new approaches to compare to the one applied in this study.

We also propose expanding the set of PII entities comprising sensitive personal data defined on LGPD, including, for example, religious and political options and PHI. Finally, studies can evaluate new topic modeling approaches by considering the sentence-level context so that the implementation can determine if a topic extracted from a new document is similar to topics discovered in the training corpus.

REFERENCES

ABBE, Adeline et al. Text mining applications in psychiatry: a systematic literature review. **International journal of methods in psychiatric research**, Wiley Online Library, v. 25, n. 2, p. 86–100, 2016.

ABDELRAZEK, Aly et al. Topic modeling algorithms and applications: A survey. **Information Systems**, Elsevier, v. 112, p. 102131, 2023.

AGGARWAL, Charu C; ZHAI, ChengXiang. **Mining text data**. [S.l.]: Springer Science & Business Media, 2012.

AGRAWAL, Amritanshu; FU, Wei; MENZIES, Tim. What is wrong with topic modeling? and how to fix it using search-based software engineering. **Information and Software Technology**, Elsevier, v. 98, p. 74–88, 2018.

AKBIK, Alan et al. FLAIR: An easy-to-use framework for state-of-the-art NLP. In: **NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)**. [S.l.: s.n.], 2019. p. 54–59.

AKBIK, Alan; BLYTHE, Duncan; VOLLGRAF, Roland. Contextual string embeddings for sequence labeling. In: **COLING 2018, 27th International Conference on Computational Linguistics**. [S.l.: s.n.], 2018. p. 1638–1649.

AKSENTIJEVIĆ, Saša; TIJAN, Edvard; AGATIĆ, Adrijana. Information security as utilization tool of enterprise information capital. In: IEEE. **2011 Proceedings of the 34th International Convention MIPRO**. [S.l.], 2011. p. 1391–1395.

ALLAHYARI, Mehdi et al. A brief survey of text mining: Classification, clustering and extraction techniques. **arXiv preprint arXiv:1707.02919**, 2017.

ALNEYADI, S.; SITHIRASENAN, E.; MUTHUKKUMARASAMY, V. Adaptable n-gram classification model for data leakage prevention. In: . [S.l.]: IEEE Computer Society, 2013. ISBN 9781479913190.

ALPAYDIN, Ethem. **Introduction to machine learning**. [S.l.]: MIT press, 2014.

ALTINEL, Berna; GANIZ, Murat Can. Semantic text classification: A survey of past and recent advances. **Information Processing & Management**, Elsevier, v. 54, n. 6, p. 1129–1153, 2018.

ALZHRANI, Khudran et al. Cnn with paragraph to multi-sequence learning for sensitive text detection. In: IEEE. **2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)**. [S.l.], 2019. p. 1–6.

____. Automated big text security classification. In: IEEE. **2016 IEEE Conference on Intelligence and Security Informatics (ISI)**. [S.l.], 2016. p. 103–108.

____. Automated big security text pruning and classification. In: IEEE. **2016 IEEE International Conference on Big Data (Big Data)**. [S.l.], 2016. p. 3629–3637.

____. Automated us diplomatic cables security classification: Topic model pruning vs. classification based on clusters. In: IEEE. **2017 IEEE International Symposium on Technologies for Homeland Security (HST)**. [S.l.], 2017. p. 1–6.

AMARAL, Daniela Oliveira Ferreira do; VIEIRA, Renata. Nerp-crf: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields. **Linguamática**, v. 6, n. 1, p. 41–49, 2014.

ANANDARAJAN, Murugan; HILL, Chelsey; NOLAN, Thomas. Text preprocessing. In: **Practical Text Analytics**. [S.l.]: Springer, 2019. p. 45–59.

ANDERSSON, Simon. Problems in information classification: insights from practice. **Information & Computer Security**, Emerald Publishing Limited, 2023.

ANGIANI, Giulio et al. A comparison between preprocessing techniques for sentiment analysis in twitter. In: **KDWeb**. [S.l.: s.n.], 2016.

ANTONS, David et al. The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. **R&D Management**, Wiley Online Library, v. 50, n. 3, p. 329–351, 2020.

ARKES, Jeremy. **Regression analysis: a practical introduction**. [S.l.]: Taylor & Francis, 2023.

ASSOCIAçãO BRASILEIRA DE NORMAS TéCNICAS. **ABNT NBR ISO/IEC 27032**: Tecnologia da informação — técnicas de segurança — diretrizes para segurança cibernética. Rio de Janeiro, Brasil, 2015. 62 p.

____. **NBR 16167**: Segurança da informação - diretrizes para classificação, rotulação, tratamento e gestão da informação. Rio de Janeiro, 2020. 16 p.

____. **ABNT NBR ISO/IEC 27002**: Segurança da informação, segurança cibernética e proteção à privacidade - controles de segurança da informação. Rio de Janeiro, Brasil, 2022. 191 p.

BELL, David Elliott; LAPADULA, Leonard J. Secure computer systems: A mathematical model, volume ii. **Journal of Computer Security**, v. 4, n. 2/3, p. 229–263, 1996.

BENDECHACHE, Malika et al. A systematic survey of data value: Models, metrics, applications and research challenges. **IEEE Access**, IEEE, 2023.

BENDECHACHE, Malika; LIMAYE, Nihar Sudhanshu; BRENNAN, Rob. Towards an automatic data value analysis method for relational databases. 2020.

BENGIO, Yoshua; GRANDVALET, Yves. Bias in estimating the variance of k-fold cross-validation. In: **Statistical modeling and analysis for complex data problems**. [S.l.]: Springer, 2005. p. 75–95.

BERGSTRÖM, Erik. **Thesis Proposal: A Method for Information Classification**. 2017.

BERGSTRÖM, Erik; ÅHLFELDT, Rose-Mharie. Information classification issues. In: SPRINGER. **Nordic Conference on Secure IT Systems**. [S.l.], 2014. p. 27–41.

BERGSTRÖM, Erik; ANTERYD, Fredrik; ÅHLFELDT, Rose-Mharie. Information classification policies: an exploratory investigation. In: INFORMATION INSTITUTE. **17th Annual Security Conference, March 26-28, 2018 Las Vegas, NV, USA**. [S.l.], 2018.

BERGSTRÖM, Erik; KARLSSON, Fredrik; ÅHLFELDT, Rose-Mharie. Developing an information classification method. **Information & Computer Security**, Emerald Publishing Limited, v. 29, n. 2, p. 209–239, 2021.

BERGSTRÖM, Erik; LUNDGREN, Martin; ERICSON, Åsa. Revisiting information security risk management challenges: a practice perspective. **Information & Computer Security**, Emerald Publishing Limited, v. 27, n. 3, p. 358–372, 2019.

BIRUNDA, S Selva; DEVI, R Kanniga. A review on word embedding techniques for text classification. **Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020**, Springer, p. 267–281, 2021.

BLEI, David M. Probabilistic topic models. **Communications of the ACM**, ACM New York, NY, USA, v. 55, n. 4, p. 77–84, 2012.

BRASIL. Lei nº 12.527, de 18 de novembro de 2011. **Diário Oficial [da] República Federativa do Brasil**, Brasília, DF, nov 2011. ISSN 1677-7042. Disponível em: <http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2011/Lei/L12527.htm>.

____. Lei nº 13.709, de 14 de agosto de 2018. **Diário Oficial [da] República Federativa do Brasil**, Brasília, DF, 2018. ISSN 1677-7042. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>.

BROWN, John David; CHARLEBOIS, Daniel. **Security classification using automated learning (scale): optimizing statistical natural language processing techniques to assign security labels to unstructured text**. [S.l.], 2010.

BUNKER, Guy. Technology is not enough: Taking a holistic view for information assurance. **Information Security Technical Report**, Elsevier, v. 17, n. 1-2, p. 19–25, 2012.

CASTRO, Pedro Vitor Quinta de; SILVA, Nádia Félix Felipe da; SOARES, Anderson da Silva. Portuguese named entity recognition using lstm-crf. In: SPRINGER. **Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13**. [S.l.], 2018. p. 83–92.

CHOWDHARY, KR1442; CHOWDHARY, KR. Natural language processing. **Fundamentals of artificial intelligence**, Springer, p. 603–649, 2020.

CHURCHILL, Rob; SINGH, Lisa. The evolution of topic modeling. **ACM Computing Surveys**, ACM New York, NY, v. 54, n. 10s, p. 1–35, 2022.

COLACE, Francesco et al. Text classification using a few labeled examples. **Computers in Human Behavior**, Elsevier, v. 30, p. 689–697, 2014.

COLLOVINI, Sandra et al. Iberlef 2019 portuguese named entity recognition and relation extraction tasks. In: **IberLEF@ SEPLN**. [S.l.: s.n.], 2019. p. 390–410.

COUCE-VIEIRA, Aitor; INSUA, David Rios; KOSGODAGAN, Alex. Assessing and forecasting cybersecurity impacts. **Decision Analysis**, INFORMS, v. 17, n. 4, p. 356–374, 2020.

DASGUPTA, Riddhiman et al. Fine grained classification of personal data entities. **arXiv preprint arXiv:1811.09368**, 2018.

DENG, Li; LIU, Yang. **Deep learning in natural language processing**. [S.l.]: Springer, 2018.

DENNING, Dorothy E. A lattice model of secure information flow. **Communications of the ACM**, ACM New York, NY, USA, v. 19, n. 5, p. 236–243, 1976.

DETROJA, Kartik; BHENSDADIA, CK; BHATT, Brijesh S. A survey on relation extraction. **Intelligent Systems with Applications**, Elsevier, p. 200244, 2023.

DIAS, Mariana et al. Named entity recognition for sensitive data discovery in portuguese. **Applied Sciences**, MDPI, v. 10, n. 7, p. 2303, 2020.

DICTIONARY, Cambridge. Intangible. In: **Cambridge English Dictionary**. [s.n.], 2021. Disponível em: <https://dictionary.cambridge.org/us/dictionary/english/intangible-asset>.

____. Tangible. In: **Cambridge English Dictionary**. [s.n.], 2021. Disponível em: <https://dictionary.cambridge.org/us/dictionary/english/tangible-asset>.

DOHERTY, Neil F; TAJUDDIN, Sharul T. Towards a user-centric theory of value-driven information security compliance. **Information Technology & People**, Emerald Publishing Limited, 2018.

DRURY, Brett; ROCHE, Mathieu. A survey of the applications of text mining for agriculture. **Computers and Electronics in Agriculture**, Elsevier, v. 163, p. 104864, 2019.

EDGAR, Thomas W; MANZ, David O. **Research methods for cyber security**. [S.l.]: Syngress, 2017.

EHRMANN, Maud et al. Named entity recognition and classification in historical documents: A survey. **ACM Comput. Surv.**, Association for Computing Machinery, New York, NY, USA, jun 2023. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3604931>.

EISENSTEIN, Jacob. **Natural language processing**. [S.l.]: MIT Press, Georgia Tech, USA, 2018.

EL-ASSADY, Mennatallah et al. Nerex: Named-entity relationship exploration in multi-party conversations. In: WILEY ONLINE LIBRARY. **Computer Graphics Forum**. [S.l.], 2017. v. 36, n. 3, p. 213–225.

ELOFF, Jan HP; HOLBEIN, Ralph; TEUFEL, Stephanie. Security classification for documents. **Computers & Security**, Elsevier, v. 15, n. 1, p. 55–71, 1996.

ELREEDY, Dina; ATIYA, Amir F. A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance. **Information Sciences**, Elsevier, v. 505, p. 32–64, 2019.

ENGELSTAD, Paal E et al. Automatic security classification with lasso. In: SPRINGER. **International Workshop on Information Security Applications**. [S.l.], 2015. p. 399–410.

____. Advanced classification lists (dirty word lists) for automatic security classification. In: IEEE. **2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery**. [S.l.], 2015. p. 44–53.

ETAIWI, Wael; NAYMAT, Ghazi. The impact of applying different preprocessing steps on review spam detection. **Procedia computer science**, Elsevier, v. 113, p. 273–279, 2017.

FARN, Kwo-Jean; LIN, Shu-Kuo; LO, Chi-Chun. A study on e-taiwan information system security classification and implementation. **Computer Standards & Interfaces**, Elsevier, v. 30, n. 1-2, p. 1–7, 2008.

FELDMAN, Ronen; SANGER, James et al. **The text mining handbook: advanced approaches in analyzing unstructured data**. [S.l.]: Cambridge university press, 2007.

FENG, Steven Y. et al. A survey of data augmentation approaches for NLP. In: **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**. Online: Association for Computational Linguistics, 2021. p. 968–988. Disponível em: <https://aclanthology.org/2021.findings-acl.84>.

FENG, Steven Y et al. A survey of data augmentation approaches for nlp. **arXiv preprint arXiv:2105.03075**, 2021.

FENZ, Stefan et al. De-identification of unstructured paper-based health records for privacy-preserving secondary use. **Journal of medical engineering & technology**, Taylor & Francis, v. 38, n. 5, p. 260–268, 2014.

FERNÁNDEZ-DELGADO, Manuel et al. An extensive experimental survey of regression methods. **Neural Networks**, Elsevier, v. 111, p. 11–34, 2019.

FERREIRA-MELLO, Rafael et al. Text mining in education. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Wiley Online Library, v. 9, n. 6, p. e1332, 2019.

FLECKENSTEIN, Mike; OBAIDI, Ali; TRYFONA, Nektaria. A review of data valuation approaches and building and scoring a data valuation model. PubPub, 2023.

FOSTER, Jonathan; CLOUGH, Paul. Embedded, added, cocreated: Revisiting the value of information in an age of data. **Journal of the Association for Information Science and Technology**, Wiley Online Library, v. 69, n. 5, p. 744–748, 2018.

GAO, Chen et al. A review on cyber security named entity recognition. **Frontiers of Information Technology & Electronic Engineering**, Springer, v. 22, n. 9, p. 1153–1168, 2021.

GARFINKEL, Steven. Executive coordination and oversight of security classification administration. **Government Information Quarterly**, Elsevier, v. 1, n. 2, p. 157–164, 1984.

GEETHA, R; KARTHIKA, S; KUMARAGURU, Ponnurangam. Tweet-scan-post: a system for analysis of sensitive private data disclosure in online social media. **Knowledge and Information Systems**, Springer, v. 63, p. 2365–2404, 2021.

GENG, Liqiang et al. Using data mining methods to predict personally identifiable information in emails. In: SPRINGER. **Advanced Data Mining and Applications: 4th International Conference, ADMA 2008, Chengdu, China, October 8-10, 2008. Proceedings 4**. [S.l.], 2008. p. 272–281.

GOEBEL, Randy et al. Explainable ai: the new 42? In: SPRINGER. **International cross-domain conference for machine learning and knowledge extraction**. [S.l.], 2018. p. 295–303.

GRECO, Marco; CRICELLI, Livio; GRIMALDI, Michele. A strategic management framework of tangible and intangible assets. **European Management Journal**, Elsevier, v. 31, n. 1, p. 55–66, 2013.

GRÖMPING, Ulrike. Variable importance in regression models. **Wiley interdisciplinary reviews: Computational statistics**, Wiley Online Library, v. 7, n. 2, p. 137–152, 2015.

GUPTA, Vishal; LEHAL, Gurpreet S et al. A survey of text mining techniques and applications. **Journal of emerging technologies in web intelligence**, v. 1, n. 1, p. 60–76, 2009.

GUTTMAN, Barbara; ROBACK, Edward A. **Sp 800-12. an introduction to computer security: the NIST handbook**. [S.l.]: National Institute of Standards & Technology, 1995.

GWET, Kilem L. **Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters**. [S.l.]: Advanced Analytics, LLC, 2014.

HACOHEN-KERNER, Yaakov; MILLER, Daniel; YIGAL, Yair. The influence of preprocessing on text classification using a bag-of-words representation. **PloS one**, Public Library of Science San Francisco, CA USA, v. 15, n. 5, p. e0232525, 2020.

HALPERIN, Morton H. Security classification and the secrecy system. **Government Information Quarterly**, Elsevier, v. 1, n. 2, p. 117–125, 1984.

HAMMER, Hugo et al. Automatic security classification by machine learning for cross-domain information exchange. In: IEEE. **MILCOM 2015-2015 IEEE Military Communications Conference**. [S.l.], 2015. p. 1590–1595.

HAPKE, Hannes Max; LANE, Hobson; HOWARD, Cole. **Natural language processing in action**. [S.l.]: Manning, 2019.

HARRISON, Suzanne; SULLIVAN, Patrick H. Profiting from intellectual capital. **Journal of intellectual capital**, MCB UP Ltd, 2000.

HARTMANN, Jochen et al. Comparing automated text classification methods. **International Journal of Research in Marketing**, Elsevier, v. 36, n. 1, p. 20–38, 2019.

HEINTZ, Ilana et al. Improving text security classification towards an automated information guard. In: IEEE. **MILCOM 2022-2022 IEEE Military Communications Conference (MILCOM)**. [S.l.], 2022. p. 757–762.

HENNESSY, SD et al. Data-centric security: Integrating data privacy and data security. **IBM Journal of Research and Development**, IBM, v. 53, n. 2, p. 2–1, 2009.

HONNIBAL, Matthew; MONTANI, Ines. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. **To appear**, v. 7, n. 1, p. 411–420, 2017.

HOWARD, Ronald A. Information value theory. **IEEE Transactions on systems science and cybernetics**, IEEE, v. 2, n. 1, p. 22–26, 1966.

HUANG, Jen-Wei; CHIANG, Chia-Wen; CHANG, Jia-Wei. Email security level classification of imbalanced data using artificial neural network: The real case in a world-leading enterprise. **Engineering Applications of Artificial Intelligence**, Elsevier, v. 75, p. 11–21, 2018.

HUANG, Zhiheng; XU, Wei; YU, Kai. Bidirectional lstm-crf models for sequence tagging. **arXiv preprint arXiv:1508.01991**, 2015.

HUYCK, C.; ORENGO, V. A stemming algorithmm for the portuguese language. In: **String Processing and Information Retrieval, International Symposium on**. Los Alamitos, CA, USA: IEEE Computer Society, 2001. p. 0186. Disponível em: <https://doi.ieeecomputersociety.org/10.1109/SPIRE.2001.10024>.

IBM Security. **Cost of a Data Breach Report 2023**. [S.l.], 2023.

IBNUGRAHA, Prajna Deshanta; NUGROHO, Lukito Edi; SANTOSA, Paulus Insap. Risk model development for information security in organization environment based on business perspectives. **International Journal of Information Security**, Springer, p. 1–14, 2020.

IGNACZAK, Luciano et al. Text mining in cybersecurity: A systematic literature review. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 54, n. 7, p. 1–36, 2021.

____. An evaluation of nerc learning-based approaches to discover personal data in brazilian portuguese documents. **Discover Data**, Springer, v. 1, n. 1, p. 5, 2023.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 27000:2018(E)**: Information technology — security techniques — information security management systems — overview and vocabulary. Geneva, Switzerland, 2018. 34 p.

JACOBI, Carina; ATTEVELDT, Wouter Van; WELBERS, Kasper. Quantitative analysis of large amounts of journalistic texts using topic modelling. **Digital Journalism**, Taylor & Francis, v. 4, n. 1, p. 89–106, 2016.

JIA, Chen; LIANG, Xiaobo; ZHANG, Yue. Cross-domain ner using cross-domain language modeling. In: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. [S.l.: s.n.], 2019. p. 2464–2474.

JO, Taeho. **Text Mining: Concepts, Implementation, and Big Data Challenge**. [S.l.]: Springer, 2018.

KAARST-BROWN, Michelle L; THOMPSON, E Dale. Cracks in the security foundation: employee judgments about information sensitivity. In: **Proceedings of the 2015 ACM SIGMIS Conference on Computers and People Research**. [S.l.: s.n.], 2015. p. 145–151.

KALETA, Jeffrey P; MAHADEVAN, Lakshman; THACKSTON, Russell. Information worth: Investigating the differences in the importance and value of personally identifiable information. **Journal of Information Systems Applied Research**, v. 16, n. 2, 2023.

KAO, Anne; POTEET, Steve R. **Natural language processing and text mining**. [S.l.]: Springer Science & Business Media, 2007.

KAPLAN, Micaela. May i ask who's calling? named entity recognition on call center transcripts for privacy law compliance. **arXiv preprint arXiv:2010.15598**, 2020.

KEISLER, Jeffrey M et al. Value of information analysis: the state of application. **Environment Systems and Decisions**, Springer, v. 34, n. 1, p. 3–23, 2014.

KITCHENHAM, Barbara Ann; BUDGEN, David; BRERETON, Pearl. **Evidence-based software engineering and systematic reviews**. [S.l.]: CRC press, 2015.

KOWSARI, Kamran et al. Text classification algorithms: A survey. **Information**, Multidisciplinary Digital Publishing Institute, v. 10, n. 4, p. 150, 2019.

KUMAR, Akshi et al. Empirical evaluation of shallow and deep classifiers for rumor detection. In: **Advances in Computing and Intelligent Systems**. [S.l.]: Springer, 2020. p. 239–252.

KUMAR, HM Keerthi; HARISH, BS. Classification of short text using various preprocessing techniques: An empirical evaluation. In: **Recent Findings in Intelligent Computing Techniques**. [S.l.]: Springer, 2018. p. 19–30.

LANDIS, J Richard; KOCH, Gary G. The measurement of observer agreement for categorical data. **biometrics**, JSTOR, p. 159–174, 1977.

LEITNER, Elena; REHM, Georg; MORENO-SCHNEIDER, Julian. Fine-grained named entity recognition in legal documents. In: SPRINGER. **Semantic Systems. The Power of AI and Knowledge Graphs: 15th International Conference, SEMANTiCS 2019, Karlsruhe, Germany, September 9–12, 2019, Proceedings**. [S.l.], 2019. p. 272–287.

LEMAÃŽTRE, Guillaume; NOGUEIRA, Fernando; ARIDAS, Christos K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. **Journal of machine learning research**, v. 18, n. 17, p. 1–5, 2017.

LI, Jing et al. A survey on deep learning for named entity recognition. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 34, n. 1, p. 50–70, 2020.

LI, Xiao-Bai; LIU, Xiaoping; MOTIWALLA, Luvai. Valuing personal data with privacy consideration. **Decision Sciences**, Wiley Online Library, v. 52, n. 2, p. 393–426, 2021.

LIANG, Yan et al. Automatic security classification based on incremental learning and similarity comparison. In: IEEE. **2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)**. [S.l.], 2019. p. 812–817.

LIU, Zihan et al. Crossner: Evaluating cross-domain named entity recognition. In: **Proceedings of the AAAI Conference on Artificial Intelligence**. [S.l.: s.n.], 2021. v. 35, n. 15, p. 13452–13460.

LOPER, Edward; BIRD, Steven. Nltk: The natural language toolkit. **arXiv preprint cs/0205028**, 2002.

LORÈ, Filippo et al. An ai framework to support decisions on gdpr compliance. **Journal of Intelligent Information Systems**, Springer, p. 1–28, 2023.

LU, Xiaofeng et al. Privacy information security classification study in internet of things. In: IEEE. **2014 International Conference on Identification, Information and Knowledge in the Internet of Things**. [S.l.], 2014. p. 162–165.

MALI, Manisha; ATIQUE, Mohammad. The relevance of preprocessing in text classification. In: SPRINGER. **Proceedings of Integrated Intelligence Enable Networks and Computing: IIENC 2020**. [S.l.], 2021. p. 553–559.

MATHKOUR, Hassan; TOUIR, Ameur; AL-SANIE, Waleed. Automatic information classifier using rhetorical structure theory. In: **Intelligent Information Processing and Web Mining**. [S.l.]: Springer, 2005. p. 229–236.

MCCALLISTER, Erika. **Guide to protecting the confidentiality of personally identifiable information**. [S.l.]: Diane Publishing, 2010.

MCHUGH, Mary L. Interrater reliability: the kappa statistic. **Biochemia medica**, Medicinska naklada, v. 22, n. 3, p. 276–282, 2012.

MILOSEVIC, Nikola; DEHGHANTANHA, Ali; CHOO, Kim-Kwang Raymond. Machine learning aided android malware classification. **Computers & Electrical Engineering**, v. 61, p. 266 – 274, 2017. ISSN 0045-7906.

MINER, Gary et al. **Practical text mining and statistical analysis for non-structured text data applications**. [S.l.]: Academic Press, 2012.

MÖLLER, Dietmar PF. **Cybersecurity in Digital Transformation Scope and Applications**. [S.l.]: Springer, 2020.

MOODY, Daniel L; WALSH, Peter. Measuring the value of information-an asset valuation approach. In: **ECIS**. [S.l.: s.n.], 1999. p. 496–512.

MORROW, Bill. Byod security challenges: control and protect your most sensitive data. **Network Security**, Elsevier, v. 2012, n. 12, p. 5–8, 2012.

MOUSTAKA, Vaia; VAKALI, Athena; ANTHOPOULOS, Leonidas G. A systematic review for smart city data analytics. **ACM Computing Surveys (CSUR)**, ACM, v. 51, n. 5, p. 103, 2018.

MUNKOVÁ, Daša; MUNK, Michal; VOZÁR, Martin. Data pre-processing evaluation for text mining: transaction/sequence model. **Procedia Computer Science**, Elsevier, v. 18, p. 1198–1207, 2013.

NADEAU, David; SEKINE, Satoshi. A survey of named entity recognition and classification. **Lingvisticae Investigationes**, John Benjamins, v. 30, n. 1, p. 3–26, 2007.

NAGPAL, Abhinav; DASGUPTA, Riddhiman; GANESAN, Balaji. Fine grained classification of personal data entities with language models. In: **5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD)**. [S.l.: s.n.], 2022. p. 130–134.

NAKAYAMA, Hiroki et al. **doccano: Text Annotation Tool for Human**. 2018. Software available from https://github.com/doccano/doccano. Disponível em: <https://github.com/doccano/doccano>.

NEERBEK, Jan; ASSENT, Ira; DOLOG, Peter. Detecting complex sensitive information via phrase structure in recursive neural networks. In: SPRINGER. **Pacific-Asia Conference on Knowledge Discovery and Data Mining**. [S.l.], 2018. p. 373–385.

NETO, Manoel Veríssimo dos Santos; SILVA, Nádia Félix F da; SOARES, Anderson da Silva. A survey and study impact of tweet sentiment analysis via transfer learning in low resource scenarios. **Language Resources and Evaluation**, Springer, p. 1–42, 2023.

NEWHOUSE, William et al. **Data Classification Concepts and Considerations for Improving Data Collection**. [S.l.], 2023.

NICOLLS, W. **Implementing company classification policy with the s/mime security label**. [S.l.], 2002.

NIEMIMAA, Elina; NIEMIMAA, Marko. Information systems security policy implementation in practice: from best practices to situated practices. **European journal of information systems**, Springer, v. 26, n. 1, p. 1–20, 2017.

OECD. Exploring the economics of personal data: a survey of methodologies for measuring monetary value. **OECD digital economy papers**, v. 220, p. 40, 2013.

OFFICE, Cabinet. **Government Security Classifications**. 2018. Online.

OLIVEIRA, Douglas Nunes de; MERSCHMANN, Luiz Henrique de Campos. Joint evaluation of preprocessing tasks with classifiers for sentiment analysis in brazilian portuguese language. **Multimedia Tools and Applications**, Springer, v. 80, p. 15391–15412, 2021.

O'MARA-EVES, Alison et al. Using text mining for study identification in systematic reviews: a systematic review of current approaches. **Systematic reviews**, Springer, v. 4, n. 1, p. 5, 2015.

ONWUBIKO, Cyril. Security issues to cloud computing. In: **Cloud Computing**. [S.l.]: Springer, 2010. p. 271–288.

ORANGE. **The future of digital trust-a european study on the nature of consumer trust and personal data,"**. 2014.

OSINSKI, Marilei et al. Methods of evaluation of intangible assets and intellectual capital. **Journal of Intellectual Capital**, Emerald Publishing Limited, 2017.

OTHMAN, Nouha; FAIZ, Rim; SMAÏLI, Kamel. Learning english and arabic question similarity with siamese neural networks in community question answering services. **Data & Knowledge Engineering**, Elsevier, v. 138, p. 101962, 2022.

PARK, Youngja et al. An experimental study on the measurement of data sensitivity. In: **Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security**. [S.l.: s.n.], 2011. p. 70–77.

____. Data classification and sensitivity estimation for critical asset discovery. **IBM Journal of Research and Development**, IBM, v. 60, n. 4, p. 2–1, 2016.

PAULSEN, Celia; BYERS, Robert. **Glossary of Key Information Security Terms**. [S.l.], 2019.

PEARSON, Cole; SELIYA, Naeem; DAVE, Rushit. Named entity recognition in unstructured medical text documents. In: IEEE. **2021 International Conference on Electrical, Computer and Energy Technologies (ICECET)**. [S.l.], 2021. p. 1–6.

PEDREGOSA, Fabian et al. Scikit-learn: Machine learning in python. **the Journal of machine Learning research**, JMLR. org, v. 12, p. 2825–2830, 2011.

PELTIER, Thomas R. Information classification. **Information systems security**, Taylor & Francis, v. 7, n. 3, p. 31–43, 1998.

PETROLINI, Michael; CAGNONI, Stefano; MORDONINI, Monica. Automatic detection of sensitive data using transformer-based classifiers. **Future Internet**, MDPI, v. 14, n. 8, p. 228, 2022.

PONEMON INSTITUTE. **2022 Global Encryption Trends Study**. [S.l.], 2022.

POYRAZ, Omer Ilker et al. Cyber assets at risk: monetary impact of us personally identifiable information mega data breaches. **The Geneva Papers on Risk and Insurance-Issues and Practice**, Springer, v. 45, p. 616–638, 2020.

PRESIDENT, US. Executive order 13526.“classified national security information memorandum.”. **Federal Register**, v. 75, 2009.

PROBST, Philipp; BOULESTEIX, Anne-Laure; BISCHL, Bernd. Tunability: Importance of hyperparameters of machine learning algorithms. **The Journal of Machine Learning Research**, JMLR. org, v. 20, n. 1, p. 1934–1965, 2019.

PRUSTY, Sashikanta; PATNAIK, Srikanta; DASH, Sujit Kumar. Skcv: Stratified k-fold cross-validation on ml classifiers for predicting cervical cancer. **Frontiers in Nanotechnology**, Frontiers Media SA, v. 4, p. 972421, 2022.

RAO, Divya; NG, Wee Keong. How much is your information worth—a method for revenue generation for your information. In: IEEE. **2015 IEEE international conference on big data (big data)**. [S.l.], 2015. p. 2320–2326.

____. A user-centric approach to pricing information. In: IEEE. **2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)**. [S.l.], 2016. p. 202–209.

REINSEL, David; GANTZ, John; RYDNING, John. Data age 2025: The evolution of data to life-critical. **Don't Focus on Big Data**, p. 2–24, 2017.

____. Data age 2025: the digitization of the world from edge to core. **IDC White Paper Doc# US44413318**, p. 1–29, 2018.

ROCHA, Naila Camila da et al. Natural language processing to extract information from portuguese-language medical records. **Data**, MDPI, v. 8, n. 1, p. 11, 2022.

RODRIGUEZ, Juan D; PEREZ, Aritz; LOZANO, Jose A. Sensitivity analysis of k-fold cross validation in prediction error estimation. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 32, n. 3, p. 569–575, 2009.

ROSS, Ron et al. **Systems Security Engineering: Cyber Resiliency Considerations for the Engineering of Trustworthy Secure Systems**. [S.l.], 2018.

SAHA, Aakanksha et al. Secrets in source code: Reducing false positives using machine learning. In: IEEE. **2020 International Conference on COMmunication Systems & NETworkS (COMSNETS)**. [S.l.], 2020. p. 168–175.

SAJKO, Mario; RABUZIN, Kornelije; BAČA, Miroslav. How to calculate information value for effective security risk assessment. **Journal of Information and Organizational Sciences**, Fakultet organizacije i informatike Sveučilišta u Zagrebu, v. 30, n. 2, p. 263–278, 2006.

SAMMUT, Claude; WEBB, Geoffrey I. **Encyclopedia of machine learning and data mining**. [S.l.]: Springer Publishing Company, Incorporated, 2017.

SEGURA-BEDMAR, Isabel; FERNÁNDEZ, Paloma Martínez; ZAZO, María Herrero. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. [S.l.], 2013.

SHA, Kewei et al. On security challenges and open issues in internet of things. **Future Generation Computer Systems**, Elsevier, v. 83, p. 326–337, 2018.

SHORTEN, Connor; KHOSHGOFTAAR, Taghi M. A survey on image data augmentation for deep learning. **Journal of big data**, SpringerOpen, v. 6, n. 1, p. 1–48, 2019.

SHORTEN, Connor; KHOSHGOFTAAR, Taghi M; FURHT, Borko. Text data augmentation for deep learning. **Journal of big Data**, Springer, v. 8, p. 1–34, 2021.

SINGH, Jasmeet; GUPTA, Vishal. A systematic review of text stemming techniques. **Artificial Intelligence Review**, Springer, v. 48, n. 2, p. 157–217, 2017.

SINGH, Kumar Pal; RISHIWAL, Vinay; KUMAR, Pramod. Classification of data to enhance data security in cloud computing. In: IEEE. **2018 3rd International Conference On Internet of Things: Smart Innovation and Usages (IoT-SIU)**. [S.l.], 2018. p. 1–5.

SIPONEN, Mikko; BASKERVILLE, Richard; KUIVALAINEN, Tapio. Integrating security into agile development methods. In: IEEE. **Proceedings of the 38th Annual Hawaii International Conference on System Sciences**. [S.l.], 2005. p. 185a–185a.

SMIRNOVA, Alisa; CUDRÉ-MAUROUX, Philippe. Relation extraction using distant supervision: A survey. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 51, n. 5, p. 1–35, 2018.

SMITH, Ray. An overview of the tesseract ocr engine. In: IEEE. **Ninth international conference on document analysis and recognition (ICDAR 2007)**. [S.l.], 2007. v. 2, p. 629–633.

SOKOLOVA, Marina; LAPALME, Guy. A systematic analysis of performance measures for classification tasks. **Information processing & management**, Elsevier, v. 45, n. 4, p. 427–437, 2009.

SOLMS, Rossouw Von; NIEKERK, Johan Van. From information security to cyber security. **computers & security**, Elsevier, v. 38, p. 97–102, 2013.

SOUZA, Ellen et al. Characterising text mining: a systematic mapping review of the portuguese language. **IET Software**, Wiley Online Library, v. 12, n. 2, p. 49–75, 2018.

SOUZA, Fábio; NOGUEIRA, Rodrigo; LOTUFO, Roberto. Portuguese named entity recognition using bert-crf. **arXiv preprint arXiv:1909.10649**, 2019.

____. Bertimbau: pretrained bert models for brazilian portuguese. In: SPRINGER. **Brazilian Conference on Intelligent Systems**. [S.l.], 2020. p. 403–417.

SOUZA, Frederico Dias; FILHO, João Baptista de Oliveira e Souza. Embedding generation for text classification of brazilian portuguese user reviews: from bag-of-words to transformers. **Neural Computing and Applications**, Springer, v. 35, n. 13, p. 9393–9406, 2023.

STATES, Organization of American. **Data Classification**. 2019. Disponível em: <https://www.oas.org/en/sms/cicte/docs/ENG-Data-Classidication.pdf>.

THORLEUCHTER, D.; POEL, D. Van den. Improved multilevel security with latent semantic indexing. **Expert Systems with Applications**, v. 39, n. 18, p. 13462–13471, dez. 2012. ISSN 09574174.

THORNTON, Grant. **Intangible Assets in a Business Combination: Identifying and valuing intangibles under IFRS 3**. [S.l.]: Grant Thornton International Ltd, 2013.

TRAN, Huy; ZDUN, Uwe et al. Systematic review of software behavioral model consistency checking. **ACM Computing Surveys (CSUR)**, ACM, v. 50, n. 2, p. 17, 2017.

UYSAL, Alper Kursat; GUNAL, Serkan. The impact of preprocessing on text classification. **Information Processing & Management**, Elsevier, v. 50, n. 1, p. 104–112, 2014.

VAJJALA, Sowmya et al. **Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems**. [S.l.]: O'Reilly Media, 2020.

VAYANSKY, Ike; KUMAR, Sathish AP. A review of topic modeling methods. **Information Systems**, Elsevier, v. 94, p. 101582, 2020.

VIDALIS, Stilianos. Calculating the value of information assets. Newport Business School Working Paper Series; Volume 1, Summer 2007 No 2-3, 2010.

WEERTS, Hilde JP; MUELLER, Andreas C; VANSCHOREN, Joaquin. Importance of tuning hyperparameters of machine learning algorithms. **arXiv preprint arXiv:2007.07588**, 2020.

WEIGUO, Fan et al. Tapping into the power of text mining. **Journal of ACM, Blacksburg**, 2005.

WHITMAN, Michael E.; MATTORD, Herbert J. **Principles of Information Security**. Boston: Cengage Learning, 2017.

WONG, Tzu-Tsung; YEH, Po-Yang. Reliable accuracy estimates from k-fold cross validation. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 32, n. 8, p. 1586–1594, 2019.

YAN, Yu; SUN, Haolin; LIU, Jie. A review and outlook for relation extraction. In: **Proceedings of the 5th International Conference on Computer Science and Application Engineering**. [S.l.: s.n.], 2021. p. 1–5.

YANG, Jianwu; CHEN, Xiaoou. A semi-structured document model for text mining. **Journal of Computer Science and Technology**, Springer, v. 17, n. 5, p. 603–610, 2002.

YANG, Min et al. Laws and regulations tell how to classify your data: A case study on higher education. **Information Processing & Management**, Elsevier, v. 60, n. 3, p. 103240, 2023.

YAYIK, Apdullah et al. Deep learning-aided automated personal data discovery and profiling. **Turkish Journal of Electrical Engineering and Computer Sciences**, v. 30, n. 1, p. 167–183, 2022.

YAZIDI, Anis et al. On the feasibility of machine learning as a tool for automatic security classification: A position paper. In: IEEE. **2016 International Conference on Computing, Networking and Communications (ICNC)**. [S.l.], 2016. p. 1–6.

ZARDARI, Munwar Ali; JUNG, Low Tang. Data security rules/regulations based classification of file data using tsf-knn algorithm. **Cluster computing**, Springer, v. 19, n. 1, p. 349–368, 2016.

ZHAI, ChengXiang; MASSUNG, Sean. **Text data management and analysis: a practical introduction to information retrieval and text mining**. [S.l.]: Association for Computing Machinery and Morgan & Claypool, 2016.

ZHAO, Rui; MAO, Kezhi. Fuzzy bag-of-words model for document representation. **IEEE transactions on fuzzy systems**, IEEE, v. 26, n. 2, p. 794–804, 2017.

ZHAO, Weizhong et al. A heuristic approach to determine an appropriate number of topics in topic modeling. In: SPRINGER. **BMC bioinformatics**. [S.l.], 2015. v. 16, p. 1–10.

ZHENG, Mingming et al. An adaptive lda optimal topic number selection method in news topic identification. **IEEE Access**, IEEE, 2023.