

UNIVERSIDADE DE PASSO FUNDO
Graduate Program in Applied Computing

Master's Thesis

**NATURAL LANGUAGE PROCESSING
FOR SENSITIVE DATA
RECOGNITION AND PRIVACY IN
DIGITAL DOCUMENTS**

SAMUEL ANTUNES VIEIRA



**UNIVERSITY OF PASSO FUNDO
INSTITUTE OF TECHNOLOGY
GRADUATE PROGRAM IN APPLIED COMPUTING**

**NATURAL LANGUAGE PROCESSING FOR
SENSITIVE DATA RECOGNITION AND
PRIVACY IN DIGITAL DOCUMENTS**

Samuel Antunes Vieira

Thesis submitted to the University of
Passo Fundo in partial fulfillment of the
requirements for the degree of Master in
Applied Computing.

Advisor: Prof. Dr. Rafael Rieder

Passo Fundo
2024

CIP – Cataloging in Publication

V658n Vieira, Samuel Antunes
 Natural language processing for sensitive data
 recognition and privacy in digital documents [electronic
 resource] / Samuel Antunes Vieira. – 2024.
 2.3 MB ; PDF.

 Advisor: Prof. Dr. Rafael Rieder.
 Thesis (Master in Applied Computing) – University of
 Passo Fundo, 2024.

 1. Data protection. 2. Automation. 3. Text classification.
 4. Digital documents. 5. Redact. I. Rieder, Rafael, advisor.
 II. Title.


 CDU: 004

Cataloging: Librarian Jucelei Rodrigues Domingues - CRB 10/1569


ATA DE DEFESA DO TRABALHO DE CONCLUSÃO DE CURSO DO ACADÊMICO

SAMUEL ANTUNES VIEIRA


Aos vinte e sete dias do mês de março do ano de dois mil e vinte e quatro, às quatorze horas, realizou-se, de forma on-line, por meio de videoconferência e presencial no Auditório da UPF ONLINE (Prédio D1), a sessão pública de defesa do Trabalho de Conclusão de Curso “NATURAL LANGUAGE PROCESSING FOR SENSITIVE DATA RECOGNITION AND PRIVACY IN DIGITAL DOCUMENTS”, de autoria de Samuel Antunes Vieira, acadêmico do Curso de Mestrado em Computação Aplicada do Programa de Pós-Graduação em Computação Aplicada – PPGCA. Segundo as informações prestadas pelo Conselho de Pós-Graduação e constantes nos arquivos da Secretaria do PPGCA, o aluno preencheu os requisitos necessários para submeter seu trabalho à avaliação. A banca examinadora foi composta pelos professores doutores Carlos Amaral Hölbig, Leonel Pablo Carvalho Tedesco e Rafael Rieder. Concluídos os trabalhos de apresentação e arguição, a banca examinadora considerou o candidato **APROVADO**. Foi concedido o prazo de até quarenta e cinco (45) dias, conforme Regimento do PPGCA, para o acadêmico apresentar ao Conselho de Pós-Graduação o trabalho em sua redação definitiva, a fim de que sejam feitos os encaminhamentos necessários à emissão do Diploma de Mestre em Computação Aplicada. Para constar, foi lavrada a presente ata, que vai assinada pelos membros da banca examinadora e pela Coordenação do PPGCA.

Documento assinado digitalmente
 **RAFAEL RIEDER**
Data: 27/03/2024 18:33:29-0300
Verifique em <https://validar.iti.gov.br>


Prof. Dr. Rafael Rieder
- UPF -
Presidente da Banca Examinadora (Orientador)

Documento assinado digitalmente
 **LEONEL PABLO CARVALHO TEDESCO**
Data: 28/03/2024 15:40:49-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Leonel Pablo Carvalho Tedesco
- UNISC -
(Avaliador Externo)

Documento assinado digitalmente
 **CARLOS AMARAL HOLBIG**
Data: 28/03/2024 09:21:36-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Carlos Amaral Hölbig
- UPF -
(Avaliador Interno)

Documento assinado digitalmente
 **CARLOS AMARAL HOLBIG**
Data: 28/03/2024 09:22:17-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Carlos Amaral Hölbig
- UPF -
(Coordenador do PPGCA)

PROCESSAMENTO DE LINGUAGEM NATURAL PARA RECONHECIMENTO E PRIVACIDADE DE DADOS CONFIDENCIAIS EM DOCUMENTOS DIGITAIS

RESUMO

Manter informações confidenciais seguras em documentos pessoais sempre foi fundamental para garantir a privacidade de pessoas ou empresas. Com a frequente digitalização de documentos e a adoção de leis e regulamentos, esta tarefa tornou-se ainda mais relevante. Neste contexto, as aplicações de segurança podem censurar textos críticos em documentos digitais. Como a proteção de dados por meio de censura pode exigir trabalho manual intensivo para identificar a localização específica de dados confidenciais e está sujeita a erros humanos, a automação é uma opção para lidar com todo o processo. Pensando nisso, este trabalho apresenta o **DOCDDOM**, um software de prova de conceito que integra múltiplas ferramentas para o reconhecimento de dados sensíveis e privacidade em documentos digitais. A abordagem considera o reconhecimento ótico de caracteres para obter dados de texto de documentos, aplica um modelo de processamento de linguagem natural focado no reconhecimento de entidades nomeadas para identificar dados confidenciais, e censura estes usando recursos de bibliotecas para processamento de documentos digitais. Os resultados preliminares mostraram que o **DOCDDOM** funciona bem, alcançando métricas de avaliação razoáveis para dois conjuntos de dados de teste de 1000 arquivos cada (Curvas AUC-PR 0,9266 e 0,6681). Uma análise detalhada identificou que existem problemas de ruído em alguns arquivos durante tarefas de classificação de texto, que ainda precisam ser tratados por meio de estratégias de distinção e filtragem de ruído. Apesar disso, a solução proposta apresentou resultados iniciais aceitáveis para uma prova de conceito, com boa precisão e acurácia para arquivos de estrutura simples e conteúdos sensíveis não numéricos.

Palavras-Chave: automação; censura; classificação de texto; documentos digitais; proteção de dados.

NATURAL LANGUAGE PROCESSING FOR SENSITIVE DATA RECOGNITION AND PRIVACY IN DIGITAL DOCUMENTS

ABSTRACT

Keeping confidential information secure in personal documents has always been critical to ensuring the privacy of people or corporations. With the frequent digitization of documents and the adoption of laws and regulations, this task has become even more relevant. In this context, security applications can redact critical texts in digital documents. As data protection through censorship can require intensive manual work to identify the specific location of sensitive data, and is subject to human error, automation is an option to handle the entire process. With this in mind, this work presents **DOCDOM**, a proof-of-concept software that integrates multiple tools for recognizing sensitive data and privacy in digital documents. The approach considers optical character recognition to obtain text data from documents, applies a natural language processing model focused on named entity recognition to identify sensitive data, and censors these using library resources for digital document processing. Preliminary results showed that **DOCDOM** works well, achieving good evaluation metrics on two test datasets of 1000 files each (AUC-PR Curves 92.66% and 66.81%). A detailed analysis identified that there are noise issues in some files during text classification tasks, which still need to be addressed through noise distinction and filtering strategies. Despite this, the proposed solution presented acceptable initial results for a proof of concept, with good precision and accuracy for files with a simple structure and sensitive non-numeric content.

Keywords: automation; data protection; digital documents; redact; text classification.

LIST OF FIGURES

Figure 1	–	Flow diagram of selection process for included studies.	13
Figure 2	–	Visual summary of selected studies: papers that contain or not content related to Personally Identifiable Information, PII (top-left); paper classification by publication type (top-right); top 5 most used acronyms (bottom-left); top 4 most used metrics (bottom-right).	17
Figure 3	–	Relationships of selected papers in terms of satisfactory solutions (y) with accuracy results (x), presence of pseudo-algorithms (color) and application on sensitive information (symbol).	18
Figure 4	–	Flask API Project Planning.	21
Figure 5	–	The flowchart illustrates the steps involved in the API workflow. The process begins with the application initialization and proceeds to processing endpoints; an instance could end with either an error or a success message. Arrows indicate the flow of control between different steps. Rhombus represents decision points. Rectangles denote both internal and external endpoints. Circles symbolize messages returned to the API clients.	25
Figure 6	–	IO Example.	26
Figure 7	–	Integration Example.	27
Figure 8	–	ROC Curve.	30
Figure 9	–	Precision-Recall Curve.	30
Figure 10	–	Comparison of Results.	31
Figure 11	–	Confusion Matrix Generated Dataset.	32
Figure 12	–	Confusion Matrix Random Dataset.	32
Figure 13	–	Random Dataset Unsuitable Output Example (p. 1).	33
Figure 14	–	Random Dataset Satisfactory Output Example (p. 1).	34

LIST OF TABLES

Table 1	–	Selected studies after SMS process.	13
Table 2	–	Summary of selected studies.	14
Table 3	–	File information organized for data input in DOCDOM	28
Table 4	–	DOCDOM output results arranged for analysis and evaluation.	28
Table 5	–	Metrics about Random Dataset Unsuitable (first line) and Satisfying (second line) outputs.	35

CONTENTS

1	INTRODUCTION	9
2	RELATED WORK	11
2.1	MAPPING REVIEW	11
2.2	MAPPING STUDY PROTOCOL	11
2.3	MAPPING STUDY RESULTS	17
3	MATERIALS AND METHODS	21
3.1	PROJECT METHODOLOGY AND AUXILIARY TOOLS	22
3.2	EVALUATION METRICS	23
3.3	DOCDOM: API WORKFLOW	24
3.4	DOCDOM: TESTING DATASET SPECIFICATIONS	26
4	RESULTS AND DISCUSSION	29
4.1	METRICS' RESULTS	29
4.2	DISCUSSION	29
4.2.1	DOCDOM process performance	29
4.2.2	Model evaluation	31
4.2.3	Noise issues	35
4.3	BENEFITS AND LIMITATIONS	36
5	CONCLUSIONS AND FUTURE WORK	38
	REFERENCES	40
	Appendix A – Results of Generated Dataset	43
	Appendix B – Results of Random Dataset	45

1. INTRODUCTION

In an epoch defined by rapid technological advancement and interconnectedness, the evolution of digital documents is like a testament to the transformative power of information technology. From humble beginnings as mere electronic facsimiles of their paper counterparts to the sophisticated, dynamic entities they are today, digital documents have revolutionized how we create, store, and disseminate information. Often, these documents encompass an array of critical content, serving as the vessel for formalizing contracts, laws, medical records, scientific papers, resumes, and many more. The ongoing digital revolution has only accelerated this reliance, a trend further exacerbated by the global pandemic outbreaks [1]. Commonly stored in portable formats such as PDF or Microsoft Word's DOCX, these files may harbor personal and private information, commonly considered sensitive data in the data science domain.

As digital documentation has taken center stage, it has triggered the emergence of a plethora of laws and regulations [2, 3, 4] designed to ensure data privacy and security. The brunt of these regulations falls on the shoulders of IT companies that develop software to manage digital data. These responsibilities extend from safeguarding personal data to fortifying information security and implementing hierarchical access controls, all aimed at averting data loss, forgery, scams, and the erosion of public trust. This panorama could lead to substantial financial losses [5]. Unfortunately, the execution of these measures is fraught with challenges. A lack of comprehensive knowledge surround information security, data leak prevention, and the handling of sensitive data has left companies exposed to the risks of sanctions, fines, and irreparable brand damage [6].

In a world where ensuring data privacy is imperative, the conventional method of manual data handling presents some challenges. Daily tasks such as identifying and redacting sensitive information, page by page, within numerous documents are labor-intensive and susceptible to human error [7]. The perils of multitasking, especially when fatigue sets in, and the influence of human emotions in jobs that require meticulous handling of sensitive data underscore the impracticality of this approach. A single oversight can have disastrous consequences, making it clear that more streamlined and accurate solutions are required.

Our research aims to address this critical problem by leveraging digital automation techniques, such as machine learning (ML), to safeguard the privacy of natural and legal persons within digital documents. The fundamental subject guiding our work is understanding the application of automation techniques to identify, classify, and accurately censor personal data, ensuring data privacy in digital documents.

The significance of this endeavor extends far beyond mere technical considerations. It is a matter of establishing and maintaining trust in digital systems. Recent legal

mandates underscore that data privacy is not a simple checkbox on an IT company's to-do list but an essential commitment to public well-being. Furthermore, our research recognizes that data privacy within digital documents is not limited to corporate policies but carries a substantial burden of time and monetary costs. The automation can offer significantly reduced costs by eliminating the human factor and the associated problems and biases from manual processes. Our research seeks to bridge the gap between theoretical insights and practical implementation, offering a more secure, efficient, and transparent approach to handling sensitive data in digital documents.

From the available automated solutions that handle confidential data, a very few have been deployed in production environments (i.e., beyond research and academia); from those, even fewer have specialized use cases with a meaningful amount of data for testings and validations. Furthermore, none of those production-like solutions have shown a professionally acceptable success ratio, such as [8] and [9]. Individual use cases might demand particular approaches, and that is where the aim of our work fits.

Furthermore, it is relevant to highlight that the companies require efficient methods for adding transparency to their data access control procedures, free from biases and discrimination [10]. Given the sensitive nature of the information involved, the corresponding documents must be stored securely and encrypted, with personal data censored or generalized as appropriate, ensuring the privacy of both consumer clients and employees.

With this in mind, this work presents **DOCDOM**, a natural language processing (NLP) proof-of-concept software for sensitive data recognition and privacy in digital documents. Our proposal applies approaches for automated document processing that focus on sensitive data identification and privacy-preserving techniques. We also consider to rigorously test and validate the accuracy of the implemented software, utilizing the appropriate metrics from real-world use cases. Through these endeavors, we seek to bridge the gap between the demand for data privacy and the practical challenges of its implementation, contributing to a more secure, efficient, and transparent approach to handling sensitive data within digital documents.

Therefore, we organized this document as follows: Section 2 presents selected studies from a systematic mapping review, showing related work; Section 3 highlights the material and methods applied for the conception of this study, detailing the **DOCDOM** solution; Section 4 presents the results obtained from a test bench, and analyses and discusses these results, pointing trends, advantages and limitations of our approach; finally, Section 5 shows conclusions and future work for continuing of this study.

2. RELATED WORK

To survey related work and gather relevant information, this study executed a systematic mapping review of the literature. The next sections present the method used, and briefly describe and discuss the selected studies.

2.1 MAPPING REVIEW

Akoka *et al.* [11] stated that a systematic mapping study (SMS) provide ways to identify research evidence, and allow the categorization and summary of context-relevant approaches. Having a written source with those features helps to maintain a progressive line that will eventually lead to the set goal.

After initial studies on the company's software development manners and overall ML concepts, a SMS was written to verify and analyse state of the art ML methods that could be used to solve the raised problems, becoming the base source of the project.

2.2 MAPPING STUDY PROTOCOL

The SMS was built having the general objective, the research questions, the research string, and the eligibility criteria as its pillars. The research covered journals and conferences published in the period of 2017-2021, aiming for the available most recent papers on the subject.

The research string considered the hole artificial intelligence (AI) field, so the results could be filtered slowly, but with increased knowledge gain in the process.

((“artificial intelligence” OR “machine learning” OR “deep learning”) AND (“document” OR “text” OR “data”) AND (“redact” OR “censor”))

The above string was adapted for each of the following databases:

- ACM: <https://dl.acm.org>;
- IEEE: <https://ieeexplore.ieee.org/Xplore/home.jsp>;
- ScienceDirect: <https://www.sciencedirect.com>;
- Springer: <https://link.springer.com>;
- Usenix: <https://www.usenix.org>.

Below, the eligibility criteria, followed by inclusion and exclusion criteria (more specific, additional filters):

1. The study must present a ML or AI approach or solution aimed at processing and/or protecting access to sensitive data in text documents
2. The study must show the application of the approach in a practical case study
3. The study must present results on the maintenance of the reliability of access to sensitive data in text documents after the use of ML or AI techniques
4. The study must include statistical analysis that highlights the accuracy of the approach, and/or a comparison with approaches related to the data redaction theme

Inclusion criteria:

- CI-01: Only studies published in journals or conferences can be selected;
- CI-02: The selected papers must be related to one of the following: computer science, engineering, machine learning, management and business.

Exclusion criteria:

- CE-01: Studies written in any language other than English will not be selected;
- CE-02: Studies whose discipline or theme (filtered in the search databases) are related to the internet of things, medicine, health, diagnostics, mobile, smart cities, facial recognition, social, harassment, 3D imaging, or Covid-19 will not be selected.

After the database search and application of eligibility criteria, the paper titles were read and duplicates removed, then paper abstracts were read, and their subject were evaluated, the ones that had very different subject with the purpose of the research were also discarded. Finally, the remaining papers were read. Figure 1 represents the full selection process. Table 1 shows the selected studies and Table 2 presents a summary of details and contributions of these studies.

The selected papers were classified as follows: a summary of the problem addressed, materials¹, methods², important acronyms³, study cases, use of sensitive data, information type, and format (e.g. type: text, format: PDF), data set volume, the training set volume, test set volume, hybrid composition⁴, algorithm usage⁵, study limitations, classification, classification justification, statistical measurements and results obtained, used metrics, and reason for the paper selection.

¹Referred to research sources, data sets, hardware, and software used by the selected studies.

²Scientific methods, architectural solutions, and algorithms used by the selected studies are considered.

³Referred to the acronyms of technologies highlighted in the selected studies, both by relevance and by usage frequency.

⁴This is the definition of whether the study used more than one solution related to machine learning algorithms or is related to other artificial intelligence techniques.

⁵It highlights whether the study contains original or scientifically known pseudo-algorithms, and if so, mentions them.

Table 1. Selected studies after SMS process.

#	Paper title
[8]	Multisource Keyword Extraction and Graph Construction for Privacy Preservation
[12]	Hybrid evolutionary approach for Devanagari handwritten numeral recognition using Convolutional Neural Network
[13]	A multi-cascaded model with data augmentation for enhanced paraphrase detection in short texts
[14]	Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering
[15]	Generalized Differential Privacy for Text Document Processing
[16]	NLP-Based Detection of Mathematics Subject Classification
[17]	Detecting Complex Sensitive Information via Phrase Structure in Recursive Neural Networks
[18]	Attention-Based Improved BLSTM-CNN for Relation Classification
[19]	A Deep Learning Model for Information Loss Prevention from Multi-Page Digital Documents

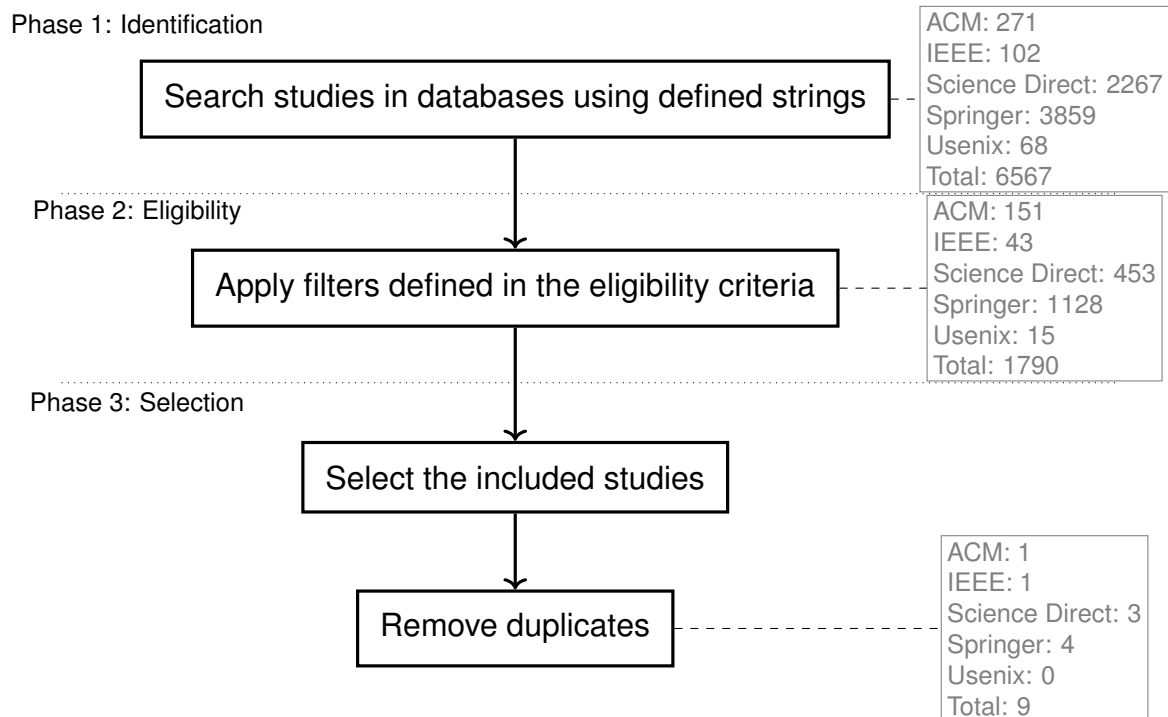


Figure 1. Flow diagram of selection process for included studies.

Table 2. Summary of selected studies.

#	Problem	Materials	Methods	Acronyms	Study Case	Contains PII	Data Format
1	For a given convoluted data containing sensitive information, the goal is to identify the instances related to the domain of the given context. Privacy preservation of these instances should be achieved by suitable transformation or encryption techniques while preserving mathematical features present in the data.	WordNet, SWCTE	Crawler, Context Detection, Semantic Graph Creation, Context Query Engine, Vector Making	PPDM, NERSO, LCS, POS, SWCTE, SEER	Using data mining for privacy preserving through keyword extraction and vector making	Yes	Text in PDF, SQL, Word
2	Although CNN achieves exceptional accuracy, still a huge number of iterations and chances to getting stuck in local optima makes it computationally expensive to train	Devanagari Numeral Dataset, Sparse Autoencoder from python SciPy, Python3, Evaluation classifier using Sklearn from python, 8GB RAM Intel Core I5 1.7GHz NVIDIA GEFORCE 820M computer	CNN, GA, L-BFGS	CNN, HWR, ReLU, GA, KNN, ECG, EMG, GAN, NN, ANN, L-BFGS	Combine 3 learning models for Devanagari handwritten numeral recognition	No	Images, handwritten numerals in database
3	Deep models have been proposed for representing and classifying paraphrases, but they require large quantities of human-labeled data, which is expensive to obtain.	MSRP, Quora, SemEval benchmark short text datasets, NetworkX, Keras, TensorFlow python libraries	Set and Graph Theory, (non)Paraphrase Generator Algorithm, BFS, DFS, Paraphrase Detector, Feature Learner, Discriminator Network,	CNN, LSTM, NLP, IR, BFS, DFS, ReLU	Paraphrase detection in short texts	No	Text in datasets
4	Text document clustering	W7 4GB RAM, Matlab R2014a	TF-IDF-FS, TF-IDF-FSDR, TF-IDF-FSDDR, LFW-FS, LFW-FSDR, LFW-FSDDR, GA, HS, PSO	GA, HS, PSO, LFW, DDR, TC, FS, TF-IDF, MAD, DDR	Text feature selection for text document clustering	No	Text
5	Author obfuscation	Fandom datasets, word2vec, fastText	Topic Classification, Earth Moving, Differential Privacy, KNN, Laplace, PDF, Gamma Distribution, PAN Obfuscation	PDF, KNN, NLP	Obfuscate authors on text documents using fandom based dataset and ready to use Application	Yes	Text
6	For authors, it can be time-consuming to find the right classification amongst thousands of choices, despite the fact that MSs labels naturally follow a hierarchical structure, and are usually presented in a sorted manner	Hardware: 24-core CPU, 64GB RAM, Software: Wolfram Language, Dataset: arXiv	Term selection, Term Vectorization, KNN and NN for digit prefixes prediction, Dimension Reduction, Optimization	MSC, NLP, KNN, SVD	Implementation of MSC classifier for 4575 MSC classes	No	Text
7	PII info detection in unstructured data relies on the frequency of co-occurrence of keywords with PII words, but this may fail to detect more complex patterns of PII	Dataset: Enron, Stanford Glove word vector set;	Complex SI Detection, RNN Training, SPR, BPTS	RNN, SPR, HIV, BPTS, MLE	Learning phrase structures that separate sensitive from non-sensitive documents in RNN	Yes	Text in documents
8	Relation Classification	Benchmark Datasets: KBP37, SemEval-2010 Task 8	Layer based: Input, Embedding, BLSTM, Attention, CNN, Output	AI, BLSTM, CNN, NLP, ML, RNN, SVM, LSTM, SDP, RCNN, PI, WV	Relation Classification using BLSTM-CNN	No	Sentences
9	Necessity of PII protection of digital documents	Dataset, n-gram tokens, RF, NB, LR, KNN, SVM	Data Collection, Sampling, Analysis, Feature Representation, Ground Truth Generation, Model Building, Re-emit	AUC, AI, ANN, BLSTM, DLP, DSS, FPR, FTC, GLBA, ITIN, INFOSEC, IPS, IDS, IPS, KNN, LDA, LR, NB, NPI, OCR, PII, PI, PCI, RF, ROC, ReLU, RNN, SVM, SD, SSN, SSE, TF-IDF, TD2V, TI, TPR, VPN, WCGM	IILPS that mines and extracts information and categorizes the document images, to SD or NSD, based on the presence of NPI and PII semantic signatures without any explicit rule configuration	Yes	Text in digital documents

Continued from previous page...

#	Source Dataset	Training Set	Test Set	Hybrid	Algorithm	Limitations	Satisfactory Solution	Satisfactory Criteria	Results Statistics	Metrics	Selection Justification
1	N/A	N/A	N/A	No	Yes: Context Graph Threshold Detection - Custom	No	No	Must implement more robust techniques or self learning algorithms instead of vector making	Average similarity score across graph 0.63	Clustering, Box Plot	Uses vector making (a field of NLP, subfield of AI) techniques for privacy preserving, it hides input context related data in text documents via censoring
2	N/A	18784	3762	Yes: CNN, GA, L-BFGS	No	No	Yes	Even though there is no algorithm or sensitive data involved, the hybrid format for handwritten numeral recognition is very useful	0.97	Accuracy, Precision, Recall, F1-score	They developed a hybrid deep learning model using GAs and L-BFGS for training CNN for handwritten numeral recognition, this paper was selected considering numbers as a part of PII
3	Quora:517968, MSRP:7814, SemEval:13231	Quora:384348, MSRP:4076, SemEval:13063	Quora:10000, MSRP:1725, SemEval:972	Yes: CNN, LSTM	Yes: Custom Data Augmentation	Yes: the impact of linguistic features is limited when the dataset is large	No	Even though there is algorithm and use of ML, paraphrase detection concept is too much away of PII recognition and redaction	Quora Accuracy 0.903, MSRP Accuracy 0.79, MSRP F1-Score 0.848, SemEval Precision 0.708, SemEval Recall 0.806, SemEval F1-Score 0.754	Accuracy, Precision, Recall, F1-score	Employs 3 supervised cascades based on CNN and LSTM for paraphrase detection, which could be possibly be used for PII detection
4	8 datasets, 22182 docs	N/A	N/A	No	Yes: DDR	No	No	Poor results	Accuracy: DS1 - 0.53, DS2 - 0.74, DS3 - 0.39, DS4 - 0.55, DS5 - 0.56, DS6 - 0.57, DS7 - 0.52, DS8 - 0.37; F-Measure: DS1 - 0.51, DS2 - 0.71, DS3 - 0.35, DS4 - 0.52, DS5 - 0.47, DS6 - 0.48, DS7 - 0.44, DS8 - 0.36	Accuracy, Dimension, F1-Score	Uses a number of feature selection methods (such as GA and KNN) for text document clustering, which could possibly determine if there is or not PII in a document
5	2 fandom based, 20, 50	N/A	N/A	No	Yes: Custom Earth Mover's Privacy Mechanism, Document Privacy Mechanism	No	No	Must explain Machine Learning techniques, not just use a already developed one (fast-Text)	Accuracy 0.937	Accuracy	They obfuscate texts by removing stylistic clues using a ML based model

Continued from previous page...

#	Source Dataset	Training Set	Test Set	Hybrid	Algorithm	Limitations	Satisfactory Solution	Satisfactory Criteria	Results Statistics	Metrics	Selection Justification
6	arXiv Bulk Data Access:4575 MSC classes	160471 papers	1000 papers	Yes: KNN, NN	No	Yes: small training data size to number of classes riation, imbalanced class representation, overlapping classes	Yes	Very detailed (though objective) combination of KNN with NN and challenges provide useful info for PII detection using NN	Recall rate: 5digit - 0.88, 3digit - 0.90, 2digit - 0.93	Recall	Combines supervised and unsupervised learning methods (KNN, NN) for a subject classification system, this combination could possibly be used to classify subjects as PII
7	Enron:1.2M+ documents	9000	Validation:1430, Test:960	No	No	Yes: supervised BPTS requires a label for each node in the tree, which is not available and would be difficult to obtain, as this would require assigning sensitivity scores to phrases of increasing complexity; complex sensitive information detection is challenging	Yes	Besides the lack of algorithm, this paper clarifies the idea of the complexity in detecting PII	Keyword-based: %ACC 0.2795 %F1 0.2004 SPRw=1: %ACC 0.3540 %F1 0.2360 SPRw=2: %ACC 0.3540 %F1 0.2400 SPRw=3: %ACC 0.7236 %F1 0.2572 SPRw=4: %ACC 0.9224 %F1 0.2536	Accuracy, F1-Score	They use RNNs to separate (unstructured) sensitive from non-sensitive documents, which is very close to our goal
8	SemEval-2010 Task 8:10717 sentences, 10 relations; KBP37:19322 sentences, 19 relations	SemEval-2010 Task 8:8000 sentences; KBP37:15917 sentences	SemEval-2010 Task 8:2717 sentences; KBP37:3405 sentences	Yes: CNN, RNN, LSTM	Yes: Custom Algorithm Procedure of Model (each Layer)	No	Yes	Even though relation classification is not exactly PII redaction, the hybrid organization of layers and hyperparameters specification shows important concepts that should be taken into consideration	CNN+WV+PF SemEval-2010 Task 8: %F1 0.789 KBP37: %F1 0.523 AI-BLSTM-CNN+WB+PF+PI SemEval-2010 Task 8: %F1 0.848 KBP37: %F1 0.637 ANN Unigram: Accuracy - 0.8879 F1-Score - 0.8735 Precision - 0.8879	F1-Score	This relation classification architecture is composed of improved BLSTM and CNN, and PII information could be detected using it
9	1.597980 Million documents	4689 documents	11723 documents	No	No	No	Yes	Although there is no algorithm, the paper highlights the performance of different methods for PII security and shows possible best solutions	Recall - 0.8606 AUC - 0.9479	Accuracy, F1-Score, Precision, Recall, AUC, True/False Positive Rate	The paper uses a number of AI methods to identify PII in texts and classify documents as SD or NSD, it does compares the results with various others methods and should be taken into account

2.3 MAPPING STUDY RESULTS

Once the necessary information was obtained, it was required to have a way of visualizing the data, facilitating the understanding of the constructed dataset as a whole, and providing material for discussion and conclusions. After the construction of the dataset, a script written in Python language with the help of data visualization libraries generated figures that could aid in the discussion of results and lead to the first conclusions. Figure 2 and Figure 3 are some examples. The first show percentages of occurrence, and the second, relation of researched projects accuracy (ascending from left to right) with presence of PII and scripts (color and format) and a custom criteria for acceptance of the solution.

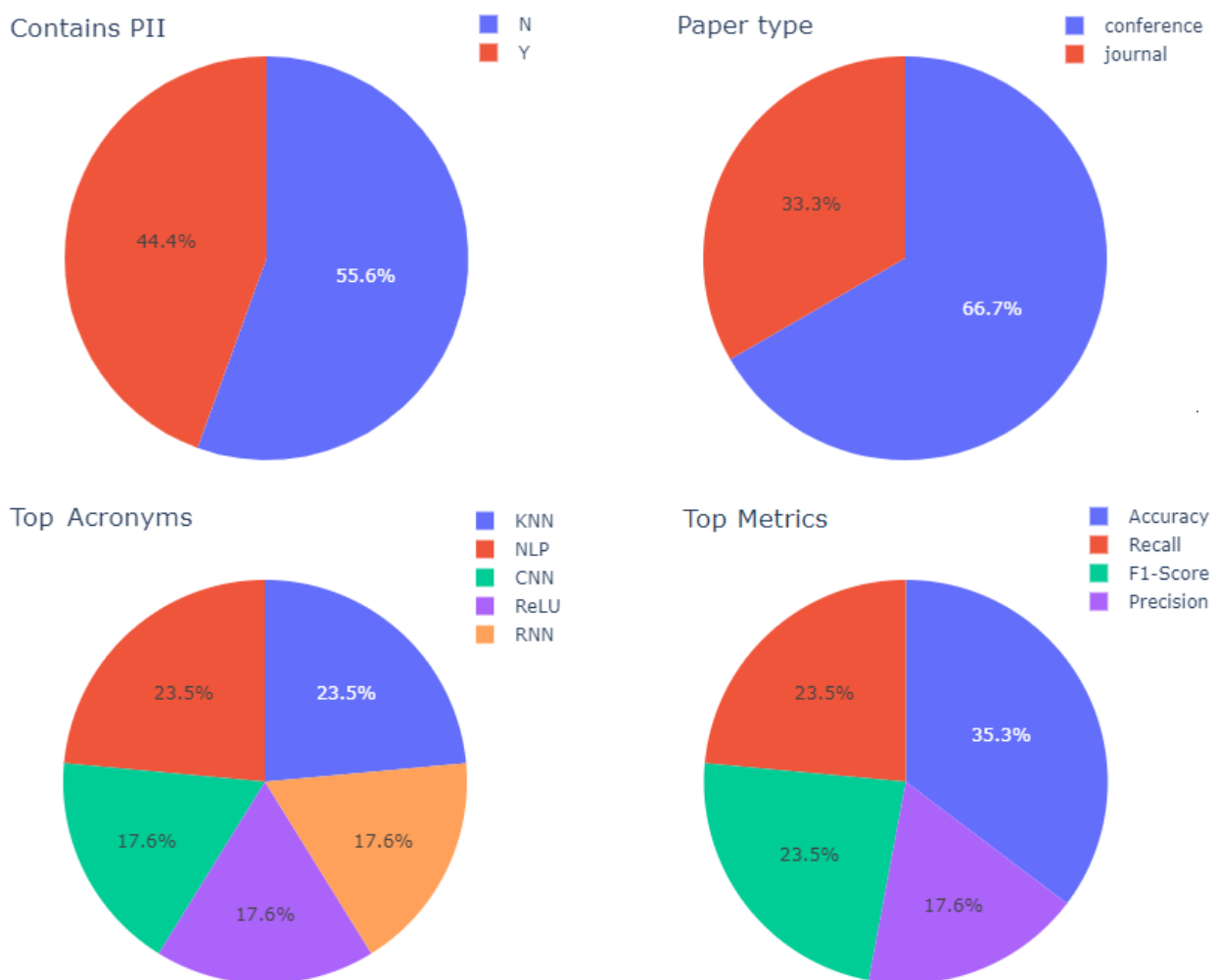


Figure 2. Visual summary of selected studies: papers that contain or not content related to Personally Identifiable Information, PII (top-left); paper classification by publication type (top-right); top 5 most used acronyms (bottom-left); top 4 most used metrics (bottom-right).

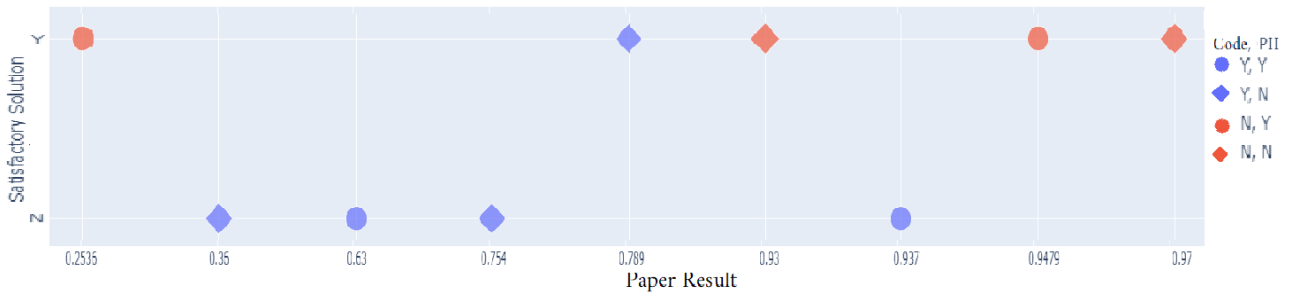


Figure 3. Relationships of selected papers in terms of satisfactory solutions (y) with accuracy results (x), presence of pseudo-algorithms (color) and application on sensitive information (symbol).

Figure 2, bottom-left, presents the most relevant acronyms found by the percentage of occurrence. Acronyms refer to ML methods used or considered in a particular study case. Figure 2, bottom-right, shows the most used metrics for the generation and analysis of the articles' results. This quantitative count is pertinent because it indicates paths to analyze results and reach better performance.

The distribution in Figure 3 allows us to observe that what classifies the selected studies as satisfactory sources of knowledge, in more weight, is the results' precision and the transparency of the process carried out in the publications over the presence of pseudo-algorithms or the specific operation with sensitive data. Thus, papers that involved relevant processes with similar objectives had significantly more utility for this study than those that claimed to have better results or omitted critical information about optional details that could lead to wrong paths of choice during implementation.

According to Figure 2, bottom-left, the most consolidated solutions for our type of problem would be one of the following: KNN (K-Nearest Neighbors), a supervised learning algorithm used for regression and classification; NLP (Natural Language Processing), an AI sub-field that seeks ways of computers to be able to understand natural languages and define linguistic concepts in text documents; CNN (Convolutional Neural Network), a deep learning subfield that focus on visual analysis; ReLU (Rectified Linear Unit), specific activation functions in artificial neural networks; and RNN (Recurrent Neural Network), another deep learning subfield, with applications such as speak recognition thanks to its memory structure [20, 21, 22].

From the selected publications, the best study cases found were the more in-depth studies, as the combination of three learning models for Devanagari handwritten numeral recognition [12], learning about sentence structures that separate confidential from non-confidential documents with RNN [17], and the creation of a system that mines and extracts information and classifies the document's images as secure or not secure based on the presence of semantic sensitive data signatures without need for explicit rule configuration [19].

The use of hybrid approaches, combining different methods like GA (Genetic Algorithm), CNN, RNN, L-BFGS (Limited-memory Broyden–Fletcher–Goldfarb–Shann), and

LSTM (Long Short-Term Memory), allow them to complement each other, generating results of higher precision. Besides, we noted relationships between results, dataset volumes, and test data types, seeming to affect the number of true positives. These relationships were not explored in any previous study, making this topic attractive for more detailed research. For results in the subject with significant precision, it is safe to say that it requires high-performance equipment (hardware) and a substantial amount of diverse documents, making it possible to sustain sufficient iterations for the learning model to generate professionally acceptable results.

Analyzing Figure 2, bottom-right, and Figure 3, it is possible to observe that the best results (considering precision above 78%) had made use of four main evaluation metrics: Accuracy, Recall, Precision, and F1-Score.

Of all the papers studied, there is no guarantee for the correct classification of data, sensitive or not. Therefore, it reinforces the development of other methods to assist this decision or confirm results with a high assurance level. In addition, in some of the reviewed studies, among others, the authors report that there is a limit, in terms of iterations, of how many times the supervised method is capable of improving the data classification, depending on the volume and the variety of content used for the model training.

Hybrid methods, composing ML techniques linked over a serial process, have generated better results. By considering specifically ML for data identification and protection, NLP has stood out for its expertise in language processing and ease of setting, performing tasks that involve both CNN and RNN approaches, and using packages and libraries that offer this assistance in various programming languages. However, by being a supervised method (that requires a lot of data to feed the network training), its hit rate is considerably dependent on the amount of learning material available. This kind of problem has alternative solutions, such as the creation of specific networks for documents of certain characteristic resemblance (where the hits are specialized in a particular type of data, making the training less expensive) and the automated generation of documents for training (adding more complexity to the system). Still, in critical cases, which aim for hit accuracy greater than 95%, the complementary implementation of particular solutions for each case is advisable, requiring an advanced knowledge of the operated project. For example, in an application that seeks to censor text fractions in documents that contain private information, one could evaluate the effectiveness of the implementation in ML and complement its flaws through censorship algorithms that previously contained the defined configuration of the location of a fraction of the sensitive information, through the standardization of processed documents.

Considering the launching of new ML approaches, the tendency is that the identification of sensitive data leads to increasingly accurate results. Regardless, currently known methods or consolidated techniques that may or may not involve hybrid systems must be improved, aiming at solutions that manipulate sensitive and non-sensitive data in accuracy levels closer to 100%.

In summary, it was realized that, besides the lack of high performance solutions for our particular problematic in the specified period, KNN and NLP were the most commonly utilized methods for word classification, and that accuracy, recall, f1-score and precision are relevant measures for the evaluation of results. Additionally, for our situation, NLP would be the best fit automation technique for developing the workflow step of processing digital documents for personal information identification, and a significant amount of training data would be required to provide acceptable results. Finally, additional methods should be used to complement more critical parts of the solution (for instance, in case the accuracy was not high enough to identify all sensitive information on a particular structure of an input document).

3. MATERIALS AND METHODS

This chapter describes stages, selected tools, processes, libraries, and other materials and methods that supported the development of the solution for privacy in digital documents, named **DOCDOM**, as well as the evaluation procedure and the documentation.

Figure 4 presents the steps for the **DOCDOM** API modeling. After the Flask project creation, we first integrated Swagger as the API interface and testing tool. The next step considered the creation of base endpoints and the integration of optical character recognition with Pytesseract. After this, we integrated natural language processing with Presidio, and the next phase was to inject the named entity recognition-focused transformer from Spacy. Subsequently, we included multiple document processing libraries for file handling and integrated Mailgun as the email service. Afterward, we gathered and generated our portable document datasets. The last step was to apply tests and evaluate the results. The following sections will discuss these steps in more detail.

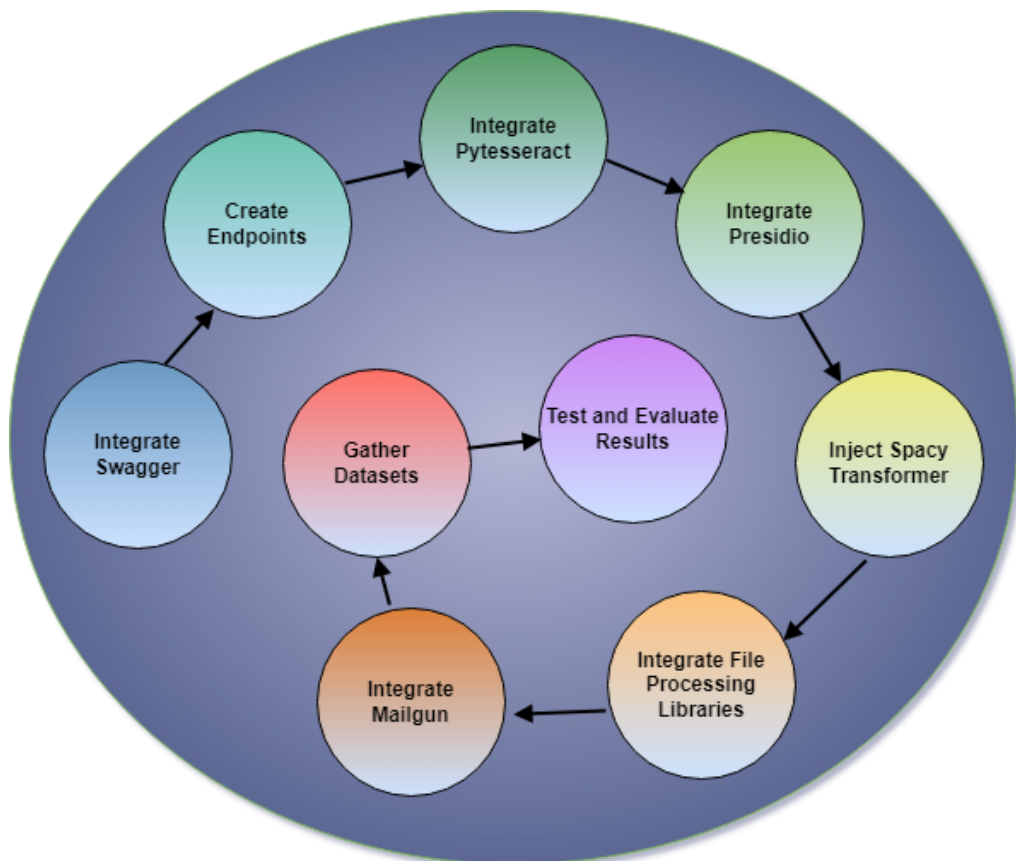


Figure 4. Flask API Project Planning.

3.1 PROJECT METHODOLOGY AND AUXILIARY TOOLS

Most consumers will work with digital documents that are either uploaded by their users or generated on their workflows. The ideal next step is to protect these recently created documents by sending them to a system that provides PII data security. The idea of our final project is that at the creation of a document, an independent API will verify the existence of personal information, identify its location, and draw black filled rectangles on the found positions, censoring PII data.

Following, if no exceptions occurred, consumers could receive the redacted version of the sent document on the informed email by our integrated solution (as one of the arguments that are sent along with the file). The consumer must then handle access control, showing only the redacted document to the appropriate parties involved in their workflow.

However, to get there, a few other treads must be handled. As the precision of which the personal data is identified depends on how well the learning model performs, it is important to 'feed' it with a considerable amount of training data. With this in mind, we opted to choose a free pre-trained model with the support of Microsoft Presidio⁶, a full documented python built "context aware, pluggable and customizable PII anonymization service for text and images" [23]. This option gave us more time for training a custom model.

Along with Presidio, we included some other functionalities, having in mind the need of Python compatibility, we opted for various libraries that could add features to our project: OpenCV and Pillow for image processing (might be required to handle the possibility of unsearchable text PDF documents, who are harder to parse text from, without some pre-processing first), Pytesseract for file conversion (as we will deal with multiple file formats) and OCR, PyPDF2 and PyMuPDF for PDF processing, and docx for document processing (additionally is able to handle rendering and rectangle drawing starting from sets of bi-dimensional points, same for the PDF libraries).

The chosen architecture for the project was Flask, being simple to manage and integrate with REST endpoints. For development assistance, documentation and local tests, Swagger and Pytest were remarkable tools.

We also injected in our project the *en_core_web_trf*, a Spacy transformer with pre-trained pipelines. Besides being free and compatible, its design structure considers handling sensitive data and achieving the highest accuracy. Additionally, this transformer was valuable for setting some hyperparameters on the NER part of the NLP pipeline of our solution.

The test phase required a dataset composed of template or curriculum vitae (CV) documents with defined PII data. For this reason, we developed a PDF file generator script to generate 1000 PDF documents containing PII data. Moreover, we collected another 1000 random PDF files (containing or not PII data) from multiple sources on the internet. These

⁶Available at <https://github.com/microsoft/presidio>

random documents have no owner and were carefully selected to ensure they had searchable text. Both generated and random file collections are available through Google Drive sharing⁷.

We manually counted and labeled every sensitive word in the 1000 random files to the base of results at section 4.1. The number of sensitive words in the 1000 generated files was always the same. The count of non-sensitive words considered the use of document editing tools, which provide the total number of words from which we subtracted the number of sensitive words previously counted.

Being one of the most known REST compatible email services, Mailgun was chosen to be integrated in our system. This way, emails of redacted files can be sent directly to third parties, allowing us to spare **DOCDOM** clients of requiring additional services to attend their customers.

Other worth to mention tools, not used in the software development, but instead that assisted in the knowledge gathering and dissemination were \LaTeX , Mendeley, DrawIO and Astah for research and design, Google Docs and Github for dataset creation and storage, and dash, numpy, plotly, pandas, seaborn and matplotlib for data visualization with Python.

3.2 EVALUATION METRICS

As discussed at section 2.3, state-of-the-art projects in similar areas usually do their tests considering Accuracy, Recall, F1-Score, and Precision evaluation metrics. Therefore, as this work also has the objective to work with a word classification type of Neural Network automation, we applied all of these, with addition of the ROC and PR curves. Following, a short explanation of each metric:

Confusion Matrix: a technique for summarizing the performance of a classification algorithm. It is a table with two rows and two columns that reports four kinds of results: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). This allows more detailed analysis than simply observing the proportion of correct classifications.

Accuracy: the degree of proximity of measurements of a numerical quantity to the actual value of that quantity.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Recall: evaluates how many of the true positives the studied model captures, when labeled as positive.

$$Recall = \frac{TP}{(TP + FN)}$$

⁷For requesting access to the file datasets, access https://drive.google.com/drive/folders/19WzP12QtvODBKPCWVY1bnovnQWdbDctC?usp=drive_link

Precision: indicates how accurately the studied model deviates from the predicted positives (how many of them are true positives).

$$Precision = \frac{TP}{(TP + FP)}$$

F1-Score: originated from Precision and Recall, this metric is the harmonic mean of both (whose maximum value is 1.0 and the minimum value is 0.0).

$$F1 - Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

Receiver Operating Characteristic (ROC) curve: shows the performance of a binary classification as a function of its cutoff threshold. It is essentially the ratio of true positives against the rate of false negatives by various threshold values. The graph x-axis represent the false positive fraction while the y-axis represent the true positive fraction.

$$x = \frac{FP}{(FP + TN)}$$

$$y = \frac{TP}{(TP + FN)}$$

Precision-Recall (PR) curve: It illustrates the trade-off between precision and recall for a binary classification model across different threshold values. Precision, representing the accuracy of positive predictions, is depicted on the y-axis, while recall, indicating the model's ability to identify all positive instances, is depicted on the x-axis. It is mathematically represented as:

$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

3.3 DOCDOM: API WORKFLOW

DOCDOM is a solution for sensitive data recognition and privacy in digital documents. It is written in Python, as a Flask based RESTful integration API. Our project contains a template generator script for the first testing dataset, as well as two main modules, one for checking the server status and another for redacting digital files.

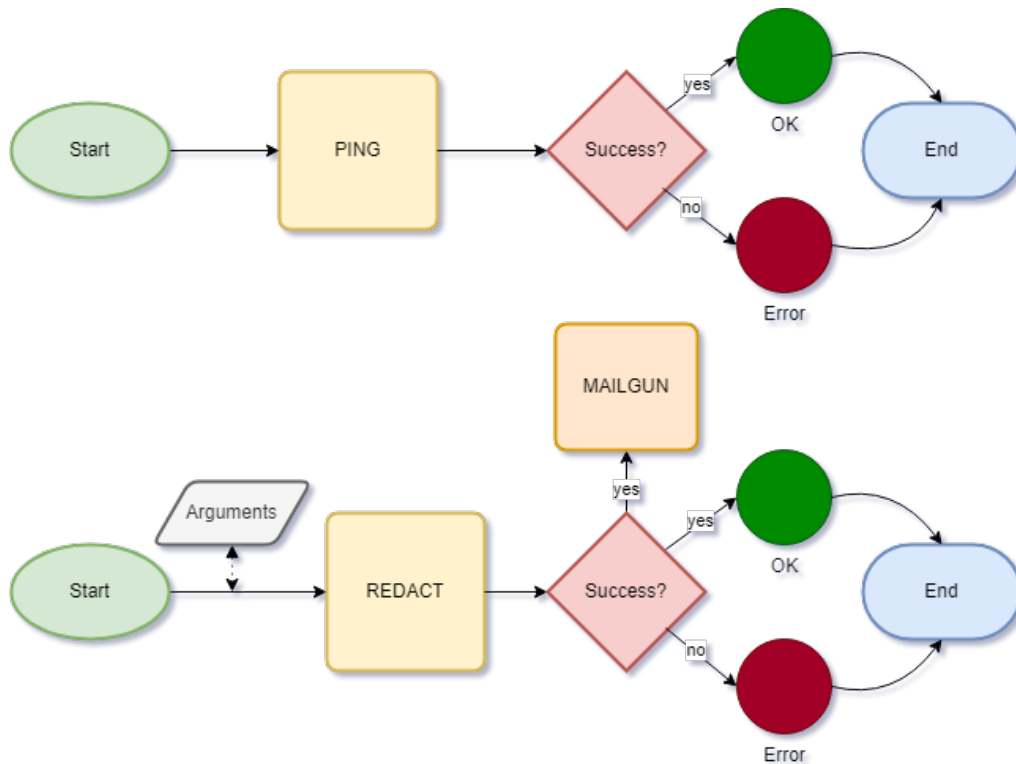


Figure 5. The flowchart illustrates the steps involved in the API workflow. The process begins with the application initialization and proceeds to processing endpoints; an instance could end with either an error or a success message. Arrows indicate the flow of control between different steps. Rhombus represents decision points. Rectangles denote both internal and external endpoints. Circles symbolize messages returned to the API clients.

Figure 5 illustrates two instances of the API processing a request (one for each available endpoint). The API runs accordingly to the subsequent flow:

1. The application is waiting for requests;
2. Requests are made either internally via Swagger or externally via REST endpoints. There are two endpoints:
 - (a) The first is PING (GET type) that checks if the server is responding;
 - (b) The second is REDACT (POST type), performing the data protection of PDF documents. This endpoint has 2 required plus 1 optional argument:
 - i. A file attachment to be redacted (required);
 - ii. A string list of one or more emails to where Mailgun will send the redacted file (required);
 - iii. A string list of keywords to increment to the list of redacted words (optional).
3. Successful requests will return an "OK" message, and unsuccessful requests will return an error. Additionally, successful POST Requests will trigger Mailgun, sending the redacted file to the email(s) informed at (ii), while unsuccessful POST requests,

for reasons such as missing arguments, corrupted files, or internal errors will return correspondent error messages and won't trigger Mailgun;

Figure 6 shows an example of the Input and Output of our system. In the POST request (file redaction), the input file should be a portable document format. If required arguments are satisfied when the file arrives in our API, we run OCR to get pure text sent to the NLP model. At this point, the words get classified as PII or not. Next, the PII words are collected and identified in the text using document processing libraries and marked into non-searchable text in black rectangles. Finally, the output file is generated and sent to Mailgun to send emails with the attached redacted file.

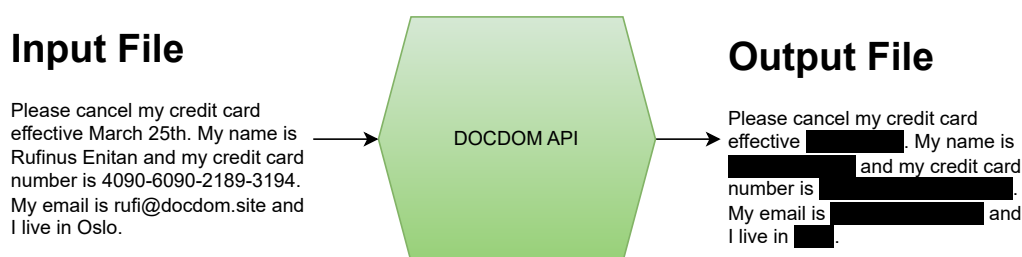


Figure 6. IO Example.

Figure 7 represents the ease of integration of our API with other solutions. Due to REST compatibility, any application can communicate with the provided endpoints. No workarounds are required since the redacting module and the Spacy Transformer are directly integrated into our system. Finally, we integrate with Mailgun independently, so the bill for this service is ours to pay. There are two envisioned possibilities regarding how third parties could deal with the redacted PDFs from the POST endpoint. The application could send the files directly to their customer's email storage or a specified department's email storage to handle the redacted files.

3.4 DOCDOM: TESTING DATASET SPECIFICATIONS

For this project, we selected two datasets for the test phase. The succeeding words represent the considered API entities for both: CREDIT_CARD, CRYPTO, DATE_TIME, EMAIL_ADDRESS, IBAN_CODE, IP_ADDRESS, LOCATION, PERSON, PHONE_NUMBER, US_BANK_NUMBER, US_PASSPORT, US_SSN. Details about each specified entity (as well as additional ones) can be found at the Microsoft's GitHub Repository⁸. These entities can be also referred as classes by the community, so for this case, we are working with a 12 class model.

To show transparency, the methodology for the results organization abide by the following set of rules:

⁸Available at https://microsoft.github.io/presidio/supported_entities/

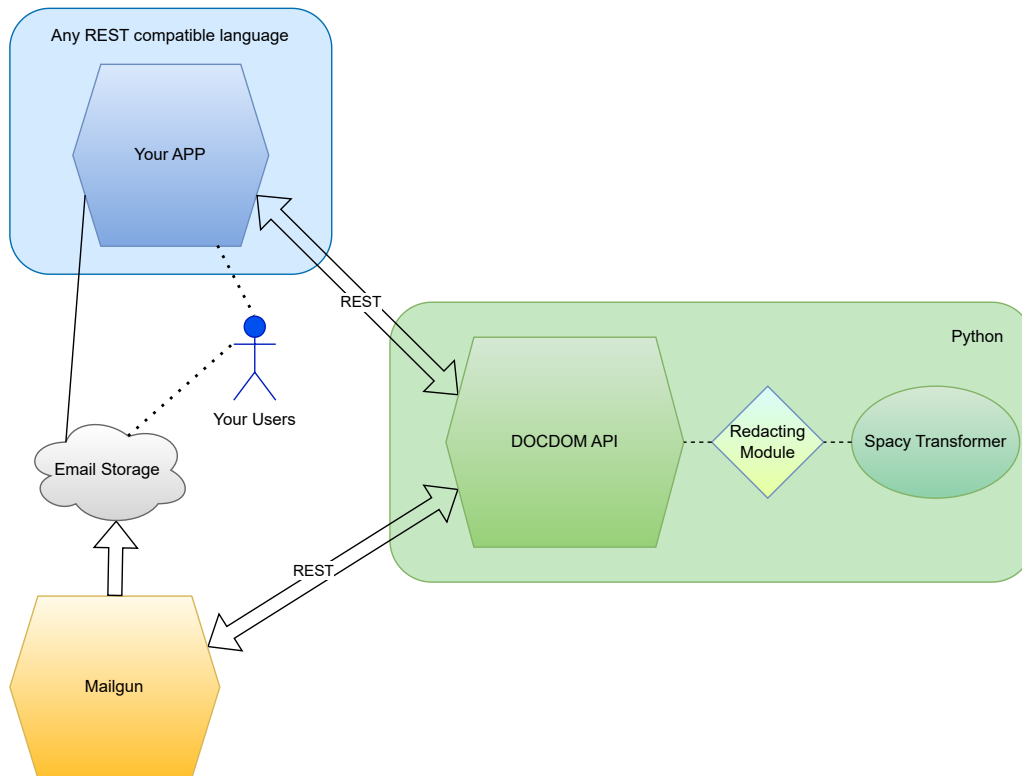


Figure 7. Integration Example.

- Date_time is considered PII only for absolute date values, meaning that day, month and year are required;
- Locations are considered PII if they are an address containing province, region or country, having part of it censored is a success;
- Company or college names are not considered FP if redacted;
- The rest of the entities are considered TP when a part or full length of these are redacted;
- The word counting process was done with the help of an online word counter software⁹ along with local file reading tools;
- Aside from word counting, every other data was either manually verified or calculated using a matching formula.

Both datasets have the same columns, divided between File information (Table 3) and Test results (Table 4). As mentioned at Section 3.1, the difference betwixt them are that the first dataset is uniquely compound of generated PDFs based on a template with little text complexity and structure diversity, and the second dataset is the opposite, accommodating various versions of documents, which are more noisy and diverse. This variation is intended

⁹Available at <https://wordcounter.net/>

4. RESULTS AND DISCUSSION

There are a few subjects we would like to elaborate on in this chapter, considering the results of evaluation metrics, as well as discussions regarding **DOCDOM** process performance, model evaluation analysis, and noise issues. Finally, we present a summary about benefits, suggestions and limitations of our solution.

4.1 METRICS' RESULTS

Appendix A and Appendix B present a part of the model evaluation results. As the results are a compound of a thousand rows (one per file) for each dataset, we chose to share them via public GitHub repository¹⁰. The spreadsheet file contains a single downloadable .xlsx format. The first page (Appendix A) has the Generated dataset results. The second page (Appendix B) has the Random dataset results.

Figure 10 presents the main results by comparing Random (blue) and Generated (red) datasets' average results. Figure 8 shows a more adequate representation of the average ROC curve for both datasets. Figure 9 shows the difference in dataset results regarding Precision and Recall measurements. Finally, Figure 11 and Figure 12 represent, respectively, the confusion matrices for the Generated dataset and the Random dataset.

4.2 DISCUSSION

The next subsections present analysis and discussion of the results, organized into three topics: system performance, model evaluation, and noise issues.

4.2.1 DOCDOM process performance

Regarding our system process execution time, we noticed that the number of files in the same output folder, as well as increased file complexity incremented the amount of processing and output generation time. File complexity time variation was linear and expected, however the time variation of processing regarding the number of files in the same output folder was apparently exponential and unexpected. The first generated files took less than two minutes to redact, whilst the last ones took more than 30 minutes.

We executed the tests in a personal computer, CPU 2.60GHz, RAM 32GB, NVIDIA GeForce GTX 1660 Ti 6GB GDDR6, with Windows 10 Home Single Language 64-bit as

¹⁰Available at https://github.com/samu158820/DOCDOM_results/blob/main/relat_samples_PII_metrics.xlsx

operating system (OS)¹¹ and fairly updated hardware. We suspect this OS architecture might be the culprit of the increasing delay for output generation. We are still in the process of thoroughly investigating this occurrence, but we do not see this particular event as critical problem in our system since it does not interfere in the actual API production environment.

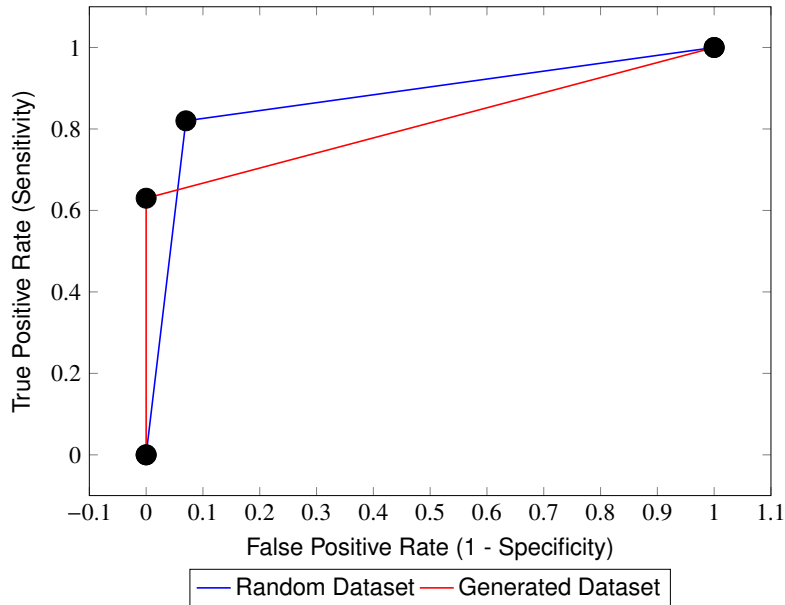


Figure 8. ROC Curve.

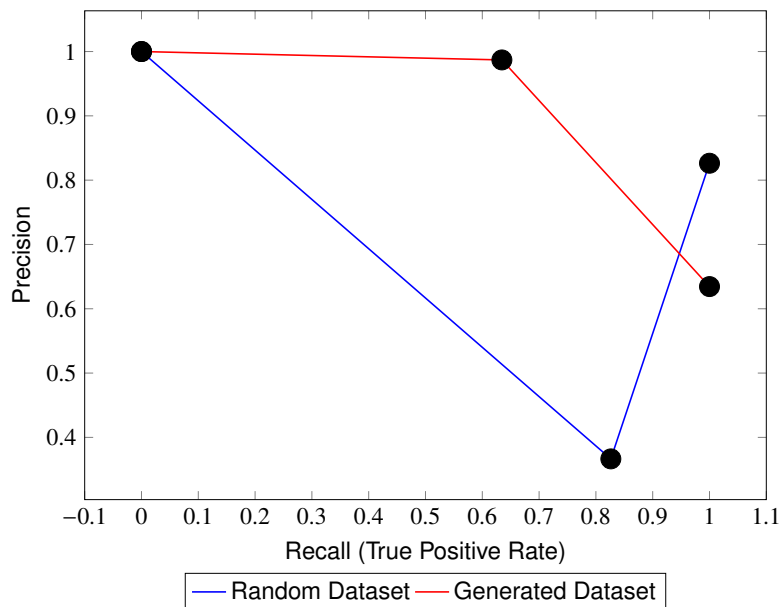


Figure 9. Precision-Recall Curve.

¹¹More details about operating system and hardware: https://github.com/samu158820/DOCDOM_results/blob/main/DxDiag.txt

4.2.2 Model evaluation

Although operating systems might affect the execution time, it shouldn't significantly impact the efficiency of internal procedures since both PyTesseract OCR and Presidio are multiplatform tools. Even before checking out figures, we could see that the generated dataset has a lowest of 5 TP and a highest of 12 TP (being 12 the same amount of PII words for all files in this dataset) with a low number of FP. The random dataset results were significantly more unstable, performing well on some cases and almost random in others. Our system performed particularly better with addresses, emails, names and dates, and operated scarcely with numbers, concluding that we must improve the regex feature.

Checking Figure 10, we can see that Accuracy and F1-score metrics had close results, in both cases the generated dataset performed better. A high Accuracy indicates that most PII and non PII have been positively identified, whereas F1-score depends on Precision and Recall. We can see that the generated dataset performed well in terms of precision, mostly due to the low number of false positives, or non-PII identified incorrectly, opposite to the instability shown in the random dataset.

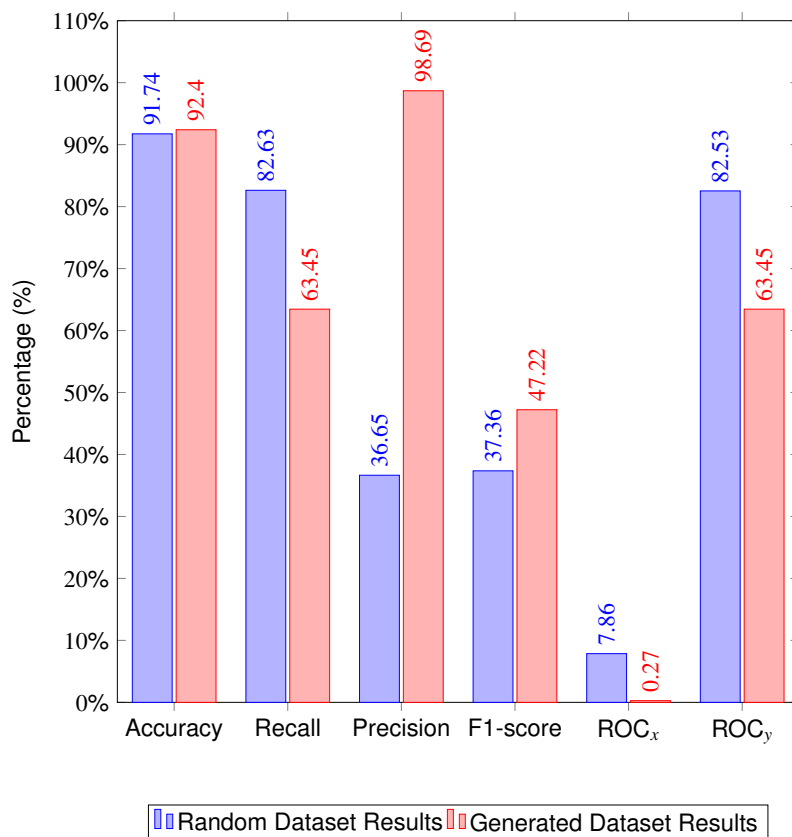


Figure 10. Comparison of Results.

Looking at Figure 11 and Figure 12, there's an overwhelming difference in the datasets between actual negatives (N = 47407 versus 559044), whereas, actual positives

(P = 12000 versus 15354) difference was not that significant. Also, by comparing positives and negatives in the same datasets (Generated: 12000P and 47407N; Random: 15354P and 559044N), we can conclude that the number of non-PII causes imbalance towards metrics such as Accuracy and Precision, which may vary more than other metrics due to their sensitivity to scale variation. This panorama means that for our case, Recall and ROC curve would be the most relevant metrics in this evaluation.

		Predicted	
		Positive	Negative
Actual	Positive	7614 (TP)	4386 (FN)
	Negative	125 (FP)	47282 (TN)

Figure 11. Confusion Matrix Generated Dataset.

		Predicted	
		Positive	Negative
Actual	Positive	12819 (TP)	2535 (FN)
	Negative	50609 (FP)	508435 (TN)

Figure 12. Confusion Matrix Random Dataset.

Usually, measurements on the same model are indifferent of which dataset is used, because there is little to none difference between the structure of datasets utilized. But for this case, as we wanted a very practical solution, with real life results, as well as different perspectives, in a way to better analyze possible defects on our system, we purposely did a comparison between documents automatically generated and random resumes from different public sources. Having this into consideration, we do not intend to compare our software with other approaches, but rather to study them to improve the **DOCDDOM** features, crossing the results and observations of the operation of our proposal.

Even though the structure complexity of the random dataset files was assumed to be one of the worst matches to our solution, it had higher TPs and lower FNs than expected, showing better Recall than the generated dataset, despite (as presumed) at the cost of much higher FPs for “noisier” files. Consequently, the calculated approximate AUC-ROC for the generated dataset was 0.815, whereas AUC-ROC for the random dataset resulted in 0.875 (calculated using the trapezoidal rule). Calculated AUC-PR curve values, in the other hand, were of 0.9266 for the generated dataset and 0.6681 for the random dataset.

CURRICULUM-VITAE

FOR: "IT UTS REVISOR"



PERMANENT ADDRESS:

[Redacted]

[Redacted], Near Al-Kabir Polytechnic College

[Redacted], O. K. K. K.

[Redacted], District - Saraikela-Kharsawan-8312

[Redacted], (Mob: 507471946 / 07 3698050)

[Redacted]-mail: [Redacted]473@gmail.com / [Redacted]91@yahoo.com

CAREER OBJECTIVE:

With an enthusiasm for achievement, I would prefer to serve an outstanding and expanding organization, which provides job satisfaction and offers me the opportunity to utilize my skills in achieving organizational goals.

ACADEMIC & PROFESSIONAL QUALIFICATION

Academic Qualification	Secondary School Certificate (SSC). In [Redacted].
Technical Qualification	Diploma in Instrumentation Engineering in [Redacted].

COMPUTER SKILLS

Diploma in Computer Application.

Knowledge about computer software.

- o Operating System : 98, XP, Vista
- o Office : Ms-Office (Word, Excel, PowerPoint)

o **Total Courses :- 11 (India/Abroad)**

Saudi Aramco - approved as an Instrument Technician.

(Ref # 7420-TRG-SA-L-144/11)

Years of Employment Records in [Redacted].

Working Period : [Redacted] February 2005.

Designation : Sr. Instrument Technician & Supervisor

Figure 13. Random Dataset Unsuitable Output Example (p. 1).

CURRICULAM-VITAE

AMIT [REDACTED]

[REDACTED], [REDACTED], [REDACTED], [REDACTED], [REDACTED]

Contact No. [REDACTED], [REDACTED]

Email:- [REDACTED]



CARRIER OBJECTIVE:-

To put my abilities and learning skills as member of team best. I would like to join a company which offers me a great chance to grow. I will work hard and try to raise the standard of your organization.

ACADEMIC QUALIFICATION:-

- Completed High School Passed from **U.P.Board** in [REDACTED].
- Completed Intermediate Passesd from **U.P.Board** in [REDACTED].
- Pursuing B.Com 2nd Year form Kanpur University.

PROFESSIONAL QUALIFICATION:-

- Basic knowledge of Computer.
- Tally.ERP9 knowledge

EXPERIENCE:-

- ❖ One Year Working Experience Aegis BPO Sector Profile As a Tele-caller.
- ❖ One Year Working Experience Airtel Call Center MNP Process Profile As a Tele-caller
- ❖ One Year Working Experience Office Of Court Counsel Profile As a Back Office

In hope to better understand the reason behind this difference between results, we analyze two files of the Random Dataset, the first, Figure 13 represents an unsuitable result, and the second, Figure 14 represents an error-free result. Table 5 compares results for these particular cases.

Table 5. Metrics about Random Dataset Unsuitable (first line) and Satisfying (second line) outputs.

Name	Color	Page	PII	nonPII	TP	FP	TN	FN	Acc	Rec	Pr	F1
160	True	5	26	818	25	23	295	1	0.37	0.96	0.04	0.08
731	False	3	10	198	10	0	198	0	1	1	1	0.5

4.2.3 Noise issues

Both the recall comparison between datasets as well as the file output comparison on the random dataset shows yet significantly different results for the same software. This situation leads us to infer that the structure of files, as well as other aspects, such as vocabulary quality, density of words and the presence of color and images have impact upon the quality of results, as for now these aspects are what we call noise. In order to achieve a better understanding of noise issues and cleaning procedures, we've selected a few relevant academic papers.

Ali [24] stated that images and colored text or background are a poor match for OCR systems and proposed that noisy text regions should be distinguished from clean text regions so that the requirement of the cleaning process should be only applied to noisy parts.

Nell [25] analyzed the noise tolerance of Tesseract OCR by experimenting with four fonts and a hundred font sizes, showing that those are another impactful noise source for text extraction.

Rotman et al. [26] present a detection network with a masking system that improves the quality of OCR execution on documents by filtering non-textual elements from the image to improve results.

Boros et al. [27] conclude that the imbalance of the datasets, the richness of the different annotation styles, and the language characteristics are important factors that might influence event detection in digitized documents.

Al et al. [28] propose a definition and taxonomy of various types of non-standard textual content (noise) in NLP to serve as a reference for researchers to consult when they devise strategies to clean and normalize noisy text.

Combining the knowledge gathered from these additional sources with the analysis of our results, our interpretation of the unstable behavior shown in the random dataset is

that our OCR model gets confused particularly by noisy files. This noise, together with clean data, is afterwards passed through the NER transformer based NLP model, which can't handle some particular cases of noisy data jointly with an unpolished regex. In summary, noise exists both in the OCR and NLP models, causing misinterpretations on the classification step and consequently reducing the quality of results.

To reduce the found noise, combining file processing packages that allow the conversion of colored documents to grayscale could improve OCR classification, and defining a minimum number of characters as well as removing punctuation characters from its input, for example, would reduce noise sent to NLP potentially enhancing its classification.

4.3 BENEFITS AND LIMITATIONS

In the market perspective, putting aside the PII redacting feature, our solution provides integration advantages such as Swagger and Mailgun. Swagger enables a very intuitive API testing system, whereas Mailgun supplies the possibility of redacted files directly and securely sent to both internal and third parties via email attachments. Additionally, if the Mailgun email feature would not be desired by customers, it could be easily turned off upon request and the redacted files would then be directly returned via async API requests.

Although our redacting software is still in its first learning steps, it already works satisfyingly well on simple structure files on most of our various listed entities mentioned in section 3.4.

For our solution, the synthesis of found problems consist of noise, in the format of structure of files, vocabulary quality, density of words, presence of color and images, variance in text font type and size, variance of annotation styles and in language characteristics.

The found results implies that in order to solve our problems, we need to improve hit rates and reduce errors, augment the model capacity for recognized entities, and set a block list for certain combinations of characters. Thereby we should seek to distinguish noisy text regions from cleaner text regions, develop additional cleaning solutions for each found noise possibility along with all non-standard textual content mentioned in [28], improve the software's regex and define strict rules for file structure.

Noise issues aside, we had a few limitations in this work. The first and foremost was the lack of staffs for labor in both the **DOCDOM** development and in the gather of results - specially to create the datasets and labeling each file. The second was dedicated time to progress further in the application of more tests and improvements that could've be done in the same period of time otherwise.

Usually IT companies work in teams for tasks that involve that much shareable labor, but we couldn't afford that much. Even disregarding research time, the system development took around a year, and the generation and gathering of results took almost six

months. During this period, we didn't find a way to fully automate the process of dataset results getting and storing data.

To give a sense of the process and the necessary amount of work from the result organization angle, we had to count all PII words in more than 2000 documents twice, manually (firstly for the preparation of results in the pre-processing time, and secondly for comparison and annotation of results in the post-processing time), which means, more than 1243610 words had to be carefully read. Fortunately, recent technology advances such as Open-AI's ChatGPT-4 [29, 30] may assist us in future API test implementations with capabilities like giving advice, summarizing, and speeding up simple labor.

5. CONCLUSIONS AND FUTURE WORK

This work presented **DOCDOM**, an NLP proof-of-concept software to automate PII security services in digital documents. Our tests show the difficulties that noise in files can induce and open room for improvement. Nevertheless, we successfully implemented a system of integrated technologies focused on a specific niche of users that require data protection in digital documents, presenting good initial results (AUC-PR Curves of 92.66% for generated dataset and 66.81% for random dataset) to identify and censor sensitive data. Once we improve our noise handling, we will launch **DOCDOM** to end-users.

Our study contributes to the community by offering an integrated data protection solution for digital documents. **DOCDOM** can already identify and redact digital pdf uncolored documents of simple structure with the most accurate results, focused on non-numeric NERs. In addition, Mailgun functionality delivers a safe transportation of attached sensitive documents, and having Swagger integrated guarantees us a well-documented API, allowing us to incorporate it in different projects easily.

As stated in section 2.3, there was a scarcity of academic papers linked to our solution, so, in addition, we built and provided two datasets. These resources contain a thousand random resume PDF files collected from different public sources and a thousand automatically generated template files using a script. We understand that these datasets can serve for new studies since the theme explored by our project is still recent in the literature and lacks fresh sources to present better results.

We also conclude that noise issues in digital documents affect both OCR and NLP models. Noise in OCR can be in the forms of file structure, presence of non-black color and images, and variance in text font and size. Whereas noise in NLP and other language models can be in the form of punctuation, vocabulary quality, variance of annotation styles due to cultural differences, and unhandled input data.

Future work in **DOCDOM** will focus on improvements in OCR by either establishing a method to distinguish between clean and noisy regions of portable documents, along with noise filters, or settling rules to accept particular file structures that would work better on our API. Supplementary, changes in NLP could reduce FPs, such as improving regex for numeric entity recognition and the appliance of rules on its output. Additionally, existing and incoming new features could use a permission system to differentiate between different levels of customer subscription plans.

One of the most time-consuming tasks of our study was to gather and fill results manually due to the lack of automation for this process. With this in mind, we will implement an automated solution for coming evaluations. After the project improvements to reduce noise aggregation, we will apply more heavy testing to select the best system configuration,

offering an improved and robust **DOCDOM** API to the market. Once up and running, we plan to publish either a documentation or a detailed product article to facilitate the adoption of new customers.

REFERENCES

- [1] MÜLLER-WOLFF, T.; HOOVEN, T. van den. Müller-wolff t, van den hooven t. arbeitshilfe zu step-up qualifizierungen und step-up personaleinsatz bei erhöhtem erkrankungsaufkommen im rahmen der sars- cov-2 herausforderungen und covid19 erkrankungen in den kliniken - empfehlungen der sektion pflegeforschung und pflegequalität. in. divi: Divi sektion pflegeforschung und pflegequalität; 2020; zugriff unter: <https://www.divi.de/empfehlungen/publikationen/covid-19/1527-divi-empfehlung-step-up-qualifizierung-pflege-covid19-2/file>. 03 2020.
- [2] COMMISSION, E. *Data protection*. [S.I.]: European Commission, 2017. https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en. Accessed: 2021-10-26.
- [3] LEGISLATION.GOV.UK. *Data Protection Act 2018*. [S.I.]: legislation.gov.uk, 2018. <https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted>. Accessed: 2021-10-26.
- [4] JUSTICE, T. U. S. D. of. *Privacy Act of 1974*. [S.I.]: The United States Department of Justice, 2021. <https://www.justice.gov/opcl/privacy-act-1974>. Accessed: 2021-10-26.
- [5] IBM. *How much does a data breach cost?* [S.I.]: IBM, 2021. <https://www.ibm.com/security/data-breach>. Accessed: 2021-11-26.
- [6] RAMAN, P.; KAYACIK, H. G.; SOMAYAJI, A. Understanding data leak prevention. In: CITESEER. *6th Annual Symposium on Information Assurance (ASIA'11)*. [S.I.], 2011. p. 27.
- [7] JEURISSEN, S. Enterprise content management: securing your sensitive data. Compact, 2020.
- [8] NETHRAVATHI, N. P. et al. Multisource keyword extraction and graph construction for privacy preservation. In: *In Proceedings of the 5th International Conference on Information and Education Technology (ICIET)*. [S.I.]: Association for Computing Machinery, 2017. p. 130–134.
- [9] PAYNE, B. Privacy protection with ai: Survey of data-anonymization techniques. 2020.
- [10] ALDBOUSH, H. H.; FERDOUS, M. Building trust in fintech: An analysis of ethical and privacy considerations in the intersection of big data, ai, and customer trust. *International Journal of Financial Studies*, MDPI, v. 11, n. 3, p. 90, 2023.
- [11] AKOKA, J.; COMYN-WATTIAU, I.; LAOUFI, N. Research on Big Data - A systematic mapping study. *Computer Standards and Interfaces*, Elsevier, v. 54, n. Part 2, p. 105 – 115, nov. 2017. Disponível em: <<https://hal.archives-ouvertes.fr/hal-01643489>>.

- [12] TRIVEDI, A. et al. Hybrid evolutionary approach for devanagari handwritten numeral recognition using convolutional neural network. *Procedia Computer Science*, v. 125, p. 525–532, 2018. ISSN 1877-0509.
- [13] SHAKEEL, M. H.; KARIM, A.; KHAN, I. A multi-cascaded model with data augmentation for enhanced paraphrase detection in short texts. *Information Processing & Management*, v. 57, n. 3, p. 102204, 2020. ISSN 0306-4573. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0306457319307502>>.
- [14] ABUALIGAH, L. M. et al. Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering. *Expert Systems with Applications*, v. 84, p. 24–36, 2017. ISSN 0957-4174. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0957417417303172>>.
- [15] FERNANDES, N.; DRAS, M.; MCIVER, A. Generalised differential privacy for text document processing. In: NIELSON, F.; SANDS, D. (Ed.). *Principles of Security and Trust*. Cham: Springer International Publishing, 2019. p. 123–148.
- [16] DONG, Y. Nlp-based detection of mathematics subject classification. In: DAVENPORT, J. H. et al. (Ed.). *Mathematical Software – ICMS 2018*. Cham: Springer International Publishing, 2018. p. 147–155.
- [17] NEERBEK, J.; ASSENT, I.; DOLOG, P. Detecting complex sensitive information via phrase structure in recursive neural networks. In: PHUNG, D. et al. (Ed.). *Advances in Knowledge Discovery and Data Mining*. Cham: Springer International Publishing, 2018. p. 373–385.
- [18] XIAO, Q. et al. Attention-based improved blstm-cnn for relation classification. In: TETKO, I. V. et al. (Ed.). *Artificial Neural Networks and Machine Learning – ICANN 2019: Text and Time Series*. Cham: Springer International Publishing, 2019. p. 34–43.
- [19] GUHA, A. et al. A deep learning model for information loss prevention from multi-page digital documents. *IEEE Access*, v. 9, p. 80451–80465, 2021.
- [20] LI, P. et al. Privacy-preserving machine learning with multiple data providers. *Future Generation Computer Systems*, v. 87, p. 341–350, 2018. ISSN 0167-739X. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0167739X17327036>>.
- [21] BOULEMTAFES, A.; DERHAB, A.; CHALLAL, Y. A review of privacy-preserving techniques for deep learning. *Neurocomputing*, v. 384, p. 21–45, 2020. ISSN 0925-2312. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0925231219316431>>.

- [22] CUNHA, M.; MENDES, R.; VILELA, J. P. A survey of privacy-preserving mechanisms for heterogeneous data types. *Computer Science Review*, v. 41, p. 100403, 2021. ISSN 1574-0137. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1574013721000435>>.
- [23] MICROSOFT. *Presidio - Data Protection and Anonymization API*. [S.l.]: GitHub, 2021. <https://github.com/microsoft/presidio>. Accessed: 2022-06-02.
- [24] ALI, M. Background noise detection and cleaning in document images. In: *Proceedings of 13th International Conference on Pattern Recognition*. [S.l.: s.n.], 1996. v. 3, p. 758–762 vol.3.
- [25] NELL, H. *Quantifying the noise tolerance of the OCR engine Tesseract using a simulated environment*. 2014.
- [26] ROTMAN, D. et al. Detection masking for improved ocr on noisy documents. *arXiv preprint arXiv:2205.08257*, 2022.
- [27] BOROS, E. et al. Assessing the impact of ocr noise on multilingual event detection over digitised documents. *International Journal on Digital Libraries*, Springer, v. 23, n. 3, p. 241–266, 2022.
- [28] SHAROU, K. A.; LI, Z.; SPECIA, L. Towards a better understanding of noise in natural language processing. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. [S.l.: s.n.], 2021. p. 53–62.
- [29] OPENAI et al. *GPT-4 Technical Report*. 2023.
- [30] OpenAI. *GPT-4*. [S.l.]: OpenAI, 2024. <https://openai.com/research/gpt-4>. Accessed: 2024-02-23.

APPENDIX A – RESULTS OF GENERATED DATASET

File Info				Results												
File_name	File_format	Colored?	Num_of_pages	Num_of_PII_words	File_label	Num_of_non_PII_words	TP	FP	TN	FN	Accuracy	Recall	Precision	F1-score	ROC_x	ROC_y
output_1	pdf	FALSE	1	46	12 TEMPLATE	46	8	0	46	4	0.931034483	0.666666667	1	0.48	0	0.666666667
output_2	pdf	FALSE	1	47	12 TEMPLATE	47	8	0	47	4	0.932203339	0.666666667	1	0.48	0	0.666666667
output_3	pdf	FALSE	1	47	12 TEMPLATE	47	7	0	47	5	0.915254237	0.583333333	1	0.465373961	0	0.583333333
output_4	pdf	FALSE	1	48	12 TEMPLATE	48	9	0	48	3	0.95	0.75	1	0.489795918	0	0.75
output_5	pdf	FALSE	1	49	12 TEMPLATE	49	8	0	49	4	0.93442623	0.666666667	1	0.48	0	0.666666667
output_6	pdf	FALSE	1	49	12 TEMPLATE	49	6	0	49	6	0.901639344	0.5	1	0.444444444	0	0.5
output_7	pdf	FALSE	1	48	12 TEMPLATE	48	9	0	48	3	0.95	0.75	1	0.489795918	0	0.75
output_8	pdf	FALSE	1	47	12 TEMPLATE	47	7	0	47	5	0.915254237	0.583333333	1	0.465373961	0	0.583333333
output_9	pdf	FALSE	1	49	12 TEMPLATE	49	8	0	49	4	0.93442623	0.666666667	1	0.48	0	0.666666667
output_10	pdf	FALSE	1	46	12 TEMPLATE	46	9	0	46	3	0.948275962	0.75	1	0.489795918	0	0.75
output_11	pdf	FALSE	1	49	12 TEMPLATE	49	8	0	49	4	0.93442623	0.666666667	1	0.48	0	0.666666667
output_12	pdf	FALSE	1	49	12 TEMPLATE	49	7	0	49	5	0.918032787	0.583333333	1	0.465373961	0	0.583333333
output_13	pdf	FALSE	1	45	12 TEMPLATE	45	10	0	45	2	0.96491281	0.833333333	1	0.495867769	0	0.833333333
output_14	pdf	FALSE	1	46	12 TEMPLATE	46	7	0	46	5	0.913793103	0.583333333	1	0.465373961	0	0.583333333
output_15	pdf	FALSE	1	47	12 TEMPLATE	47	7	0	47	5	0.915254237	0.583333333	1	0.465373961	0	0.583333333
output_16	pdf	FALSE	1	46	12 TEMPLATE	46	7	0	46	5	0.913793103	0.583333333	1	0.465373961	0	0.583333333
output_17	pdf	FALSE	1	49	12 TEMPLATE	49	10	0	49	2	0.967213115	0.833333333	1	0.495867769	0	0.833333333
output_18	pdf	FALSE	1	47	12 TEMPLATE	47	9	0	47	3	0.949152542	0.75	1	0.489795918	0	0.75
output_19	pdf	FALSE	1	49	12 TEMPLATE	49	9	0	49	3	0.950819672	0.75	1	0.489795918	0	0.75
output_20	pdf	FALSE	1	49	12 TEMPLATE	49	8	0	49	4	0.93442623	0.666666667	1	0.48	0	0.666666667
output_880	pdf	FALSE	1	45	12 TEMPLATE	45	7	0	45	5	0.912280702	0.583333333	1	0.465373961	0	0.583333333
output_881	pdf	FALSE	1	49	12 TEMPLATE	49	7	0	49	5	0.918032787	0.583333333	1	0.465373961	0	0.583333333
output_882	pdf	FALSE	1	48	12 TEMPLATE	48	9	0	48	3	0.95	0.75	1	0.489795918	0	0.75
output_883	pdf	FALSE	1	48	12 TEMPLATE	48	8	0	48	4	0.933333333	0.666666667	1	0.48	0	0.666666667
output_884	pdf	FALSE	1	48	12 TEMPLATE	48	7	0	48	5	0.916666667	0.583333333	1	0.465373961	0	0.583333333
output_885	pdf	FALSE	1	50	12 TEMPLATE	50	6	2	48	6	0.870967742	0.5	0.75	0.48	0.04	0.5
output_886	pdf	FALSE	1	46	12 TEMPLATE	46	8	0	46	4	0.931034483	0.666666667	1	0.48	0	0.666666667
output_887	pdf	FALSE	1	49	12 TEMPLATE	49	8	0	49	4	0.93442623	0.666666667	1	0.48	0	0.666666667
output_888	pdf	FALSE	1	48	12 TEMPLATE	48	7	0	48	5	0.916666667	0.583333333	1	0.465373961	0	0.583333333
output_889	pdf	FALSE	1	48	12 TEMPLATE	48	7	0	48	5	0.916666667	0.583333333	1	0.465373961	0	0.583333333
output_890	pdf	FALSE	1	49	12 TEMPLATE	49	8	0	49	4	0.93442623	0.666666667	1	0.48	0	0.666666667
output_891	pdf	FALSE	1	48	12 TEMPLATE	48	8	0	48	4	0.933333333	0.666666667	1	0.48	0	0.666666667
output_892	pdf	FALSE	1	46	12 TEMPLATE	46	9	0	46	3	0.948275962	0.75	1	0.489795918	0	0.75
output_893	pdf	FALSE	1	47	12 TEMPLATE	47	9	0	47	3	0.949152542	0.75	1	0.489795918	0	0.75
output_894	pdf	FALSE	1	47	12 TEMPLATE	47	7	0	47	5	0.915254237	0.583333333	1	0.465373961	0	0.583333333
output_895	pdf	FALSE	1	46	12 TEMPLATE	46	7	0	46	5	0.913793103	0.583333333	1	0.465373961	0	0.583333333
output_896	pdf	FALSE	1	50	12 TEMPLATE	50	7	0	50	5	0.919354839	0.583333333	1	0.465373961	0	0.583333333
output_897	pdf	FALSE	1	45	12 TEMPLATE	45	9	0	45	3	0.947368421	0.75	1	0.489795918	0	0.75
output_898	pdf	FALSE	1	46	12 TEMPLATE	46	8	0	46	4	0.931034483	0.666666667	1	0.48	0	0.666666667
output_899	pdf	FALSE	1	47	12 TEMPLATE	47	8	1	46	4	0.915254237	0.666666667	0.888888889	0.489795918	0.021276596	0.666666667
output_1000	pdf	FALSE	1	49	12 TEMPLATE	49	7	0	49	5	0.918032787	0.583333333	1	0.465373961	0	0.583333333
Average			1	47.407	12 TEMPLATE	47.407	7.614	0.125	47.282	4.386	0.924044159	0.6345	0.88694246	0.47217591	0.002650932	0.6345

APPENDIX B – RESULTS OF RANDOM DATASET

File Info					Results											
File_name	File_format	Colored?	Num_of_pages	Num_of_Pii_words	File_label	Num_of_non_Pii_words	TP	FP	TN	FN	Accuracy	Recall	Precision	F1-score	ROC_x	ROC_y
d_1	pdf	FALSE	2	9	CV	496	9	8	488	0	0.984158416	0.529411765	0.45262722	0.016129032	1	1
d_2	pdf	FALSE	2	12	CV	248	7	17	231	5	0.915384615	0.583333333	0.291696667	0.444444444	0.068548387	0.583333333
d_3	pdf	FALSE	4	16	CV	704	12	143	561	4	0.795833333	0.75	0.077419255	0.169624842	0.203125	0.75
d_4	pdf	TRUE	4	12	CV	519	8	50	469	4	0.888305085	0.666666667	0.137931034	0.284816333	0.096339114	0.666666667
d_5	pdf	TRUE	4	9	CV	724	9	7	717	0	0.990450205	1	0.5625	0.4608	0.009668508	1
d_6	pdf	TRUE	2	14	CV	456	12	37	419	2	0.917021277	0.857142857	0.244897959	0.345679012	0.08140351	0.857142857
d_7	pdf	FALSE	3	9	CV	301	6	10	291	3	0.958064516	0.666666667	0.375	0.4608	0.033222591	0.666666667
d_8	pdf	TRUE	3	9	CV	234	7	17	217	2	0.9218107	0.777777778	0.291696667	0.396894215	0.072649573	0.777777778
d_9	pdf	FALSE	2	9	CV	420	9	31	389	0	0.927738928	1	0.225	0.299875052	0.073809524	1
d_10	pdf	TRUE	2	9	CV	241	8	13	228	1	0.944	0.888888889	0.380952381	0.42	0.052941909	0.888888889
d_11	pdf	TRUE	1	9	CV	515	9	109	406	0	0.791894733	1	0.076271166	0.131689253	0.211650485	1
d_12	pdf	TRUE	3	16	CV	660	16	80	580	0	0.881656805	1	0.166666667	0.244897959	0.121212121	1
d_13	pdf	FALSE	2	10	CV	381	5	9	372	5	0.926701571	0.5	0.357142857	0.486111111	0.049723757	0.5
d_14	pdf	TRUE	4	15	CV	1502	15	272	1230	0	0.820689748	1	0.052264808	0.09403754	0.181091877	1
d_15	pdf	TRUE	3	10	CV	808	9	40	768	1	0.949877751	0.9	0.183673469	0.281528296	0.04950495	0.9
d_16	pdf	TRUE	3	16	CV	622	9	64	558	7	0.88714734	0.5625	0.123287671	0.24912259	0.102893891	0.5625
d_17	pdf	TRUE	5	9	CV	1072	9	120	952	0	0.88891674	1	0.069767442	0.121928166	0.119402299	1
d_18	pdf	TRUE	5	9	CV	681	9	31	650	0	0.955072464	1	0.225	0.299875052	0.045621252	1
d_19	pdf	FALSE	2	9	CV	316	8	5	311	1	0.981538462	0.888888889	0.615384615	0.483471074	0.015822785	0.888888889
d_20	pdf	TRUE	4	12	CV	973	12	38	935	0	0.96142132	1	0.24	0.312174616	0.039054471	1
d_980	pdf	FALSE	3	8	CV	645	8	9	636	0	0.986217458	1	0.47058235	0.4352	0.013953488	1
d_981	pdf	TRUE	3	17	CV	528	14	12	516	3	0.972477064	0.823529412	0.538461538	0.478096268	0.022727273	0.823529412
d_982	pdf	TRUE	5	21	CV	1693	21	171	1522	0	0.900233372	1	0.109375	0.177742511	0.101004135	1
d_983	pdf	FALSE	2	8	CV	161	6	10	151	2	0.928994083	0.75	0.375	0.444444444	0.062111801	0.75
d_984	pdf	TRUE	4	23	CV	1188	22	377	811	1	0.687861272	0.956521739	0.055137845	0.103063723	0.317340067	0.956521739
d_985	pdf	TRUE	2	14	CV	706	7	13	693	7	0.972222222	0.5	0.35	0.484420666	0.018413598	0.5
d_986	pdf	TRUE	3	15	CV	616	12	20	596	3	0.963549921	0.8	0.375	0.434585785	0.02467532	0.8
d_987	pdf	TRUE	5	29	CV	1175	21	110	1065	8	0.901993355	0.724137931	0.160395344	0.296796975	0.093611021	0.724137931
d_988	pdf	FALSE	1	8	CV	120	6	4	116	2	0.953125	0.75	0.6	0.49382716	0.033333333	0.75
d_989	pdf	TRUE	4	14	CV	1114	11	86	1028	3	0.921092991	0.785714286	0.113402062	0.220436653	0.077199282	0.785714286
d_990	pdf	TRUE	4	17	CV	687	12	13	674	5	0.974431818	0.705882353	0.48	0.48185941	0.019822853	0.705882353
d_991	pdf	FALSE	2	9	CV	197	9	12	185	0	0.941747573	1	0.428571429	0.42	0.069913706	1
d_992	pdf	FALSE	5	22	CV	1091	18	125	966	4	0.884097035	0.818181818	0.125874126	0.231111111	0.114573786	0.818181818
d_993	pdf	TRUE	3	19	CV	593	18	50	543	1	0.916666667	0.947368421	0.264705882	0.341392522	0.084317032	0.947368421
d_994	pdf	TRUE	3	9	CV	451	2	16	435	7	0.95	0.222222222	0.111111111	0.444444444	0.035476718	0.222222222
d_995	pdf	TRUE	2	11	CV	241	9	18	223	2	0.920634921	0.818181818	0.333333333	0.411357341	0.074688797	0.818181818
d_996	pdf	FALSE	3	16	CV	621	12	12	609	4	0.974882261	0.75	0.5	0.48	0.018323671	0.75
d_997	pdf	TRUE	2	10	CV	215	6	17	198	4	0.905666667	0.6	0.280899565	0.422405577	0.079059767	0.6
d_998	pdf	TRUE	2	14	CV	200	9	14	186	5	0.911214953	0.642857143	0.391304348	0.470416362	0.07	0.642857143
d_999	pdf	TRUE	3	14	CV	333	11	56	277	3	0.829971182	0.785714286	0.164179104	0.285932023	0.168168168	0.785714286
d_1000	pdf	FALSE	2	17	CV	236	15	18	218	2	0.920948617	0.882352941	0.454545455	0.4488	0.076271186	0.882352941
Average			2.931	15.354		559.044	12.819	50.609	508.435	2.535	0.917467572	0.826310201	0.365663358	0.373641949	0.078654241	0.825310201



UPF

UNIVERSIDADE
DE PASSO FUNDO

UPF Campus I - BR 285, São José
Passo Fundo - RS - CEP: 99052-900
(54) 3316 7000 - www.upf.br