Pseudonymization in Legal Texts according to the LGPD: A Named Entity Recognition Approach

Marcelo Anselmo[1][0009−0005−1145−3501] and Bruno César Ribas[2][0000−0001−6314−1511]

[1] Department of Computer Science, University of Brasilia, Brasilia, Brazil
marcelo.anselmo@aluno.unb.br
[2] Department of Computer Science, University of Brasilia, Brasilia, Brazil
bruno.ribas@unb.br

Abstract. This study explores the application of Named Entity Recognition (NER) for the pseudonymization of data in legal texts, aiming to protect Personally Identifiable Information (PII) in compliance with Brazil's General Data Protection Law (LGPD). The research highlights the challenge of balancing data privacy and utility, presenting a methodology that uses artificial intelligence technologies to effectively identify and obscure PII in legal documents. In this study, we propose a Transformer model along with Regex techniques to identify entities in a text. To test the model, we created a new dataset from the existing LenerBR. We also used a function and prompt engineering applied to the Llama 8B version 3 model to generate synthetic data. Tests showed the need for further adjustments in the proposed new model. Future work will focus on improving the model's accuracy and efficiency, as well as enhancing the identification of sensitive data and learning from user interactions.

Keywords: Data Pseudonymization · Named Entity Recognition · Information Privacy · Legal Texts · LGPD · Transformer.

1 Introduction

In today's digital age, where technology and computing power have grown exponentially, personal data has become an extremely valuable asset. Companies capture this information from online interactions, such as websites and social networks, to create detailed user profiles. These profiles range from consumer habits to sensitive information like health conditions and personal preferences. This practice not only allows for the personalization of online experiences, but it also exposes users to privacy risks, including the illegal sale of such data in dark markets, potentially for use in criminal activities [3].

This scenario highlights the critical importance of protecting personal data, not only for individual privacy, but also for the overall security of society. The solution to these issues has manifested through regulations and laws dedicated to data protection, such as the General Data Protection Regulation (GDPR) in the European Union and the General Data Protection Law (LGPD) in Brazil [9].

The main problem addressed in this work is the difficulty of balancing individual privacy with the need to use their information in legal and administrative texts. This issue deepens when dealing with identifying and hiding PII without compromising data integrity and utility. Using resources, such as Transformer models and Regex techniques, an effective balance between privacy and utility is sought, as already accomplished in other studies, such as the work of G. M. GR, S. Abhi, and R. Agarwal [5].

The aim of this study is to investigate and demonstrate the effectiveness of Named Entity Recognition (NER) applied to PII pseudonymization in legal documents. It proposes to develop and test a model capable of efficiently identifying and hiding named entities that correspond to PII. The Figure 1 illustrates the pipeline adopted in our methodology. Another important note is the CNMP Resolution No. 281, dated December 12, 2023 [3], in its Article 79, presents pseudonymization as an alternative to protecting the personal data of natural persons in procedures or processes within the Brazilian Public Prosecutor's Office.

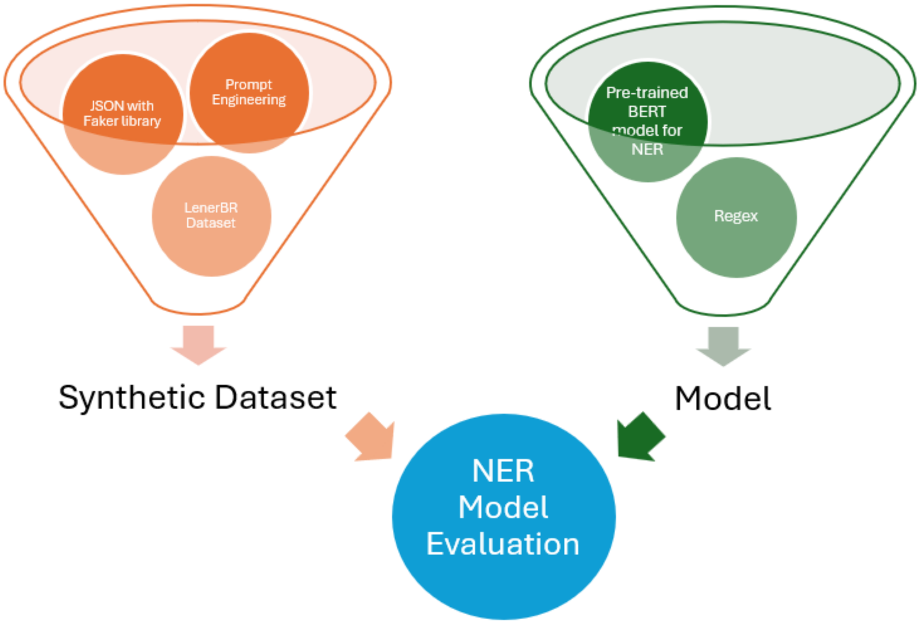


Fig. 1. Pipeline adopted in our methodology

Legal texts are often lengthy and complex, prioritizing formality over readability [12]. However, the implementation of advanced techniques, such as ar-

[3] https://www.cnmp.mp.br/portal/images/CALJ/resolucoes/Resoluo-281-de-2023.pdf

tificial intelligence models, often relies on large volumes of annotated data for training, which can be challenging in terms of cost and time [13]. With this in mind, we face a fundamental challenge: the need for data to be reliably and efficiently labeled. Traditionally, this labeling is done manually, a process that, despite improving the accuracy of artificial intelligence models, is notoriously time-consuming and expensive [4]. This scenario imposes a high cost in terms of time, money, and effort, limiting the scalability of solutions.

This paper is structured as follows: Section 2 presents the background concepts, covering fundamental concepts, like the related to personal data protection and pseudonymization. Section 3 describes related work, highlighting previous studies that addressed similar issues. Section 4 presents the proposed methodology, detailing the data used, tools, and terms to be found. Section 5 discusses the expected results for each test set. Finally, Section 6 presents the conclusions and suggestions for future work.

2 Background Concepts

2.1 General Data Protection Law (LGPD)

The General Data Protection Law (LGPD[4], Law No. 13,709, dated August 14, 2018) of Brazil is a regulation that establishes guidelines for the collection, use, processing, and storage of personal data. The law grants individuals more control over their personal information while balancing privacy rights with technological advancement. The LGPD requires that any operation involving personal data be based on clear legal justifications and respect strict principles, such as purpose, adequacy, and necessity [3].

2.2 Personal Data (PD)

Personal data (LGPD, Art. 5, I) refers to information that identifies or can directly or indirectly identify a natural person. This definition is broad and includes a variety of types of information, ranging from names and digital identities to physical and electronic characteristics, such as photos and location data. Such data is crucial in digital society, being used in various contexts, from identity verification to service personalization. However, handling it involves significant privacy risks, which requires robust protection measures to ensure data security and confidentiality [5].

2.3 Anonymization vs. Pseudonymization

Anonymization and pseudonymization are two fundamental concepts in data privacy management. Anonymization (LGPD, Art. 5, XI) refers to the process

[4] https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

[5] https://www.serpro.gov.br/lgpd/menu/protecao-de-dados/dados-pessoais-lgpd

whereby personal data is processed to remove the ability to permanently associate this data with a specific individual, without using additional information. This means that once anonymized, the data cannot be linked back to the owner, even with the use of reasonable technical means. This process is irreversible, ensuring a high level of protection for the individual's identity.

On the other hand, pseudonymization (LGPD, Art. 13, § 4) is a technique that modifies personal data so that identification of the owner cannot be done without using additional information, which is maintained separately and under strict security measures by the data controller. Unlike anonymization, pseudonymization is reversible, provided that the additional information is made available [6]. This method will be adopted in this work as it allows the data to be protected without losing its utility for analysis and research.

2.4 Retrieval-Augmented Generation (RAG)

RAG is an advanced approach in natural language processing (NLP) models that integrates information retrieval with text generation [6]. This technique allows the model to retrieve relevant documents that are used to inform subsequent text generation. The unique ability of RAG to combine parametric and non-parametric memory enables it to adapt both retrieval and content generation for specific NLP tasks, improving the accuracy and relevance of the generated information [7]. By utilizing RAG, we can incorporate a vast amount of pre-existing legal information to enhance the quality of our generated dataset.

3 Related Work

In the article by Patsakis and Lykousas (2023) [11], the authors explore the challenge of effectively anonymizing text in the age of large language models. This study questions the effectiveness of these approaches and assesses their ability to prevent identification, especially with the use of AI on large datasets. An experiment is conducted using GPT on anonymized texts of well-known personalities to verify whether these language models can re-identify individuals. They argue that despite technological advances, there are still significant obstacles in protecting data privacy in texts processed by these tools.

In the work of de Andrade et al. (2023) [1], the authors present an approach called PromptNER, which focuses on NER in sensitive data using automatically labeled instances. The authors propose an approach that employs LLMs to identify entities in complaints and then trains simpler models like the SpERT method. The improved NER model shows substantial improvements in F-score, ranging from 41% to 129% compared to models using only manually labeled data. This study is crucial because it combines artificial intelligence methods to improve efficiency in identifying and handling personal data.

[6] https://www.cnmp.mp.br/portal/images/CALJ/resolucoes/Resoluo-281-de-2023.pdf

[7] https://huggingface.co/transformers/model_doc/rag.html

Mota et al. (2021) [7] investigate the use of neural networks for named entity recognition in legal documents in Portuguese. The authors used Spacy and FLAIR libraries. Their results demonstrate the ability of these advanced technologies to handle the linguistic complexity of Portuguese, providing a technical foundation for the development of more effective solutions.

Nunes and dos Santos (2023) [9] discuss the application and impact of the LGPD in the Brazilian context, focusing on a technological and agnostic approach. They emphasize the need for organizations to adapt to legal requirements and the importance of implementing effective measures for protecting personal data. This study highlights that implementing the LGPD requires a comprehensive technical and organizational approach. Multiples tools, such as Data Management Systems (DMS), anonymization, pseudonymization, encryption, and auditing, are highlighted as essential for compliance. Techniques like Big Data and Machine Learning can improve compliance, while privacy by design and eDiscovery are emphasized as crucial for protecting data.

GR, Shinu, and Agarwal (2023) [5] presented a hybrid model that combines RegEx and NER for resume analysis and matching. This work illustrates the applicability of NER techniques along with regular expressions to efficiently extract and process information in structured documents like resumes. The proposed methodology is relevant to our study as it demonstrates the effectiveness of integrating NER and RegEx in data identification and pseudonymization tasks.

Luz de Araujo et al. (2018) [2] developed LeNER-BR, a dataset specifically for Named Entity Recognition in Brazilian legal texts. Their results provide an essential foundation for validating NER models in the Brazilian legal context. The dataset they proposed serves as a valuable resource for training and testing new methodologies, including the approach of our study.

Oliveira et al. (2024) [10] explored the combination of prompt-based language models and weak supervision to label the task of Named Entity Recognition (NER) in legal documents. They highlighted the effectiveness of their techniques in improving the accuracy and automation of the NER process, which is crucial for data pseudonymization in compliance with privacy regulations.

4 Methodology

4.1 Personal Data Used

We categorized terms according to some types of Personal Data (PD). It is important to note that the spectrum of terms is much broader than those discussed in this study, however, we highlight some popular terms. Note that the CPF is a Brazilian ID number. The terms used to generate synthetic data are presented in Table 1 below.

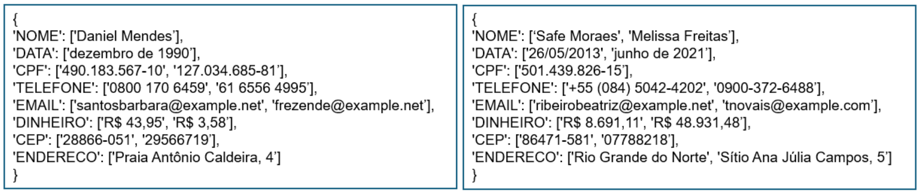
Table 1. Selected Terms for Identification of Personal Data (PD)

Term	**Example**
NAME	*João da Silva*
DATE	*12 de janeiro de 2013*
ADDRESS	*Rua do Alecrim, 0*
CPF	*123.456.789-00*
PHONE	*(11) 99999-9999*
EMAIL	*example@example.com*
MONEY	*R$ 1,000.00*
ZIP CODE	*12345-678*

4.2 Dataset

For this study, we used the LenerBR dataset derived from the study by Luz de Araujo and Pedro Henrique [2] and available at [8]. This dataset consists exclusively of legal documents. It includes labels for people, places, temporal entities, and organizations, as well as specific tags for legal entities and judicial processes.

We used the functions of the Faker library [9] to generate synthetic data. The *Faker* library is a Python tool that generates fake data as per the function used. It is useful for creating synthetic data for testing and prototyping. We created a Python function that generates random data for one to two items per term using the *Faker* library. Figure 2 below show examples of data generated by our function.



```
{
'NOME': ['Daniel Mendes’],
'DATA': ['dezembro de 1990’],
'CPF': ['490.183.567-10', '127.034.685-81’],
'TELEFONE': ['0800 170 6459', '61 6556 4995’],
'EMAIL': ['santosbarbara@example.net', 'frezende@example.net’],
'DINHEIRO': ['R$ 43,95', 'R$ 3,58'],
'CEP': ['28866-051', '29566719’],
'ENDERECO': ['Praia Antônio Caldeira, 4’]
}
```

```
{
'NOME': [‘Safe Moraes', 'Melissa Freitas’],
'DATA': ['26/05/2013', 'junho de 2021’],
'CPF': ['501.439.826-15’],
'TELEFONE': ['+55 (084) 5042-4202', '0900-372-6488’],
'EMAIL': ['ribeirobeatriz@example.net', 'tnovais@example.com’],
'DINHEIRO': ['R$ 8.691,11', 'R$ 48.931,48’],
'CEP': ['86471-581', '07788218'],
'ENDERECO': ['Rio Grande do Norte', 'Sítio Ana Júlia Campos, 5’]
}
```

Fig. 2. Example of data generated by running our function twice.

The entities in portuguese and its respective translation to english, hidden the entities that has the same translation, are: "*NOME*" (NAME), "*DATA*" (DATE), "*ENDERECO*" (ADDRESS), "*TELEFONE*" (PHONE), "*DINHEIRO*" (MONEY), "*CEP*" (ZIP CODE).

As seen in the previous images, we generated synthetic data for the term CPF. For this term, being a personal identifier, we changed the values of the check digit. According to the study by [8], the check digit is a number calculated

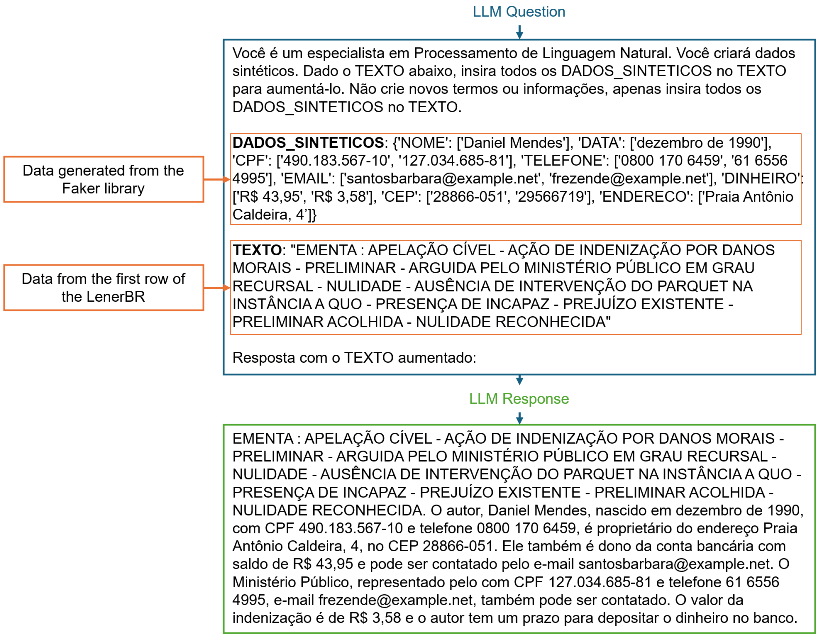
[8] `https://huggingface.co/datasets/lener_br`

[9] `https://faker.readthedocs.io/en/master/`

from the other numbers of the CPF. It is used to verify whether the number is mathematically consistent with the standards defined by the competent Brazilian agencies. This action aims to ensure that invalid CPF is generated, thus preventing the exposure of sensitive data.

We applied prompt engineering techniques with RAG to create synthetic data. For this purpose, we used the Llama model version 3 with 8 billion parameters, a large-scale open-source language model from Meta. It is freely available for research purposes [10].

Some of the most relevant parameters for running the Llama model were: "*temperature=0.2*" and "*max_new_tokens=512*". The testing environment was a local machine with a GPU of 24GB VRAM and an Intel I9 10th generation processor. For each generated text, the time was around 12 seconds. These parameters were used to generate the synthetic data. Our code is available at a github repository [11]. An example of a prompt to generate a synthetic text and the response to this prompt is shown in Figure 3 below.


LLM Question
Data generated from the Faker library
Data from the first row of the LenerBR
LLM Response

Você é um especialista em Processamento de Linguagem Natural. Você criará dados sintéticos. Dado o TEXTO abaixo, insira todos os DADOS_SINTETICOS no TEXTO para aumentá-lo. Não crie novos termos ou informações, apenas insira todos os DADOS_SINTETICOS no TEXTO.

DADOS_SINTETICOS: {'NOME': ['Daniel Mendes'], 'DATA': ['dezembro de 1990'], 'CPF': ['490.183.567-10', '127.034.685-81'], 'TELEFONE': ['0800 170 6459', '61 6556 4995'], 'EMAIL': ['santosbarbara@example.net', 'frezende@example.net'], 'DINHEIRO': ['R$ 43,95', 'R$ 3,58'], 'CEP': ['28866-051', '29566719'], 'ENDERECO': ['Praia Antônio Caldeira, 4’]}

TEXTO: "EMENTA : APELAÇÃO CÍVEL - AÇÃO DE INDENIZAÇÃO POR DANOS MORAIS - PRELIMINAR - ARGUIDA PELO MINISTÉRIO PÚBLICO EM GRAU RECURSAL - NULIDADE - AUSÊNCIA DE INTERVENÇÃO DO PARQUET NA INSTÂNCIA A QUO - PRESENÇA DE INCAPAZ - PREJUÍZO EXISTENTE - PRELIMINAR ACOLHIDA - NULIDADE RECONHECIDA"

Resposta com o TEXTO aumentado:

EMENTA : APELAÇÃO CÍVEL - AÇÃO DE INDENIZAÇÃO POR DANOS MORAIS - PRELIMINAR - ARGUIDA PELO MINISTÉRIO PÚBLICO EM GRAU RECURSAL - NULIDADE - AUSÊNCIA DE INTERVENÇÃO DO PARQUET NA INSTÂNCIA A QUO - PRESENÇA DE INCAPAZ - PREJUÍZO EXISTENTE - PRELIMINAR ACOLHIDA - NULIDADE RECONHECIDA. O autor, Daniel Mendes, nascido em dezembro de 1990, com CPF 490.183.567-10 e telefone 0800 170 6459, é proprietário do endereço Praia Antônio Caldeira, 4, no CEP 28866-051. Ele também é dono da conta bancária com saldo de R$ 43,95 e pode ser contatado pelo e-mail santosbarbara@example.net. O Ministério Público, representado pelo com CPF 127.034.685-81 e telefone 61 6556 4995, e-mail frezende@example.net, também pode ser contatado. O valor da indenização é de R$ 3,58 e o autor tem um prazo para depositar o dinheiro no banco.

Fig. 3. Prompt Engineering to generate a synthetic example.

Using this prompt model from Figure 3, we managed to generate 600 synthetic data of legal texts. These examples were used to evaluate the new proposed

[10] `https://llama.meta.com/llama3`

[11] `https://github.com/celiudos/paper_bracis_2024`

NER model for LGPD. We noticed that, even with the command passed to the prompt to use all synthetic data, the model did not use the value "*29566719*" for "*CEP*". Therefore, we created a function that removes these data from the generated JSON already displayed in 2.

4.3 NER Model for Evaluation

We used a pre-trained BERT model for named entity recognition in Portuguese, provided by Pierre [12]. This model leveraged the same dataset previously mentioned, the LenerBR. However, our study changed the names of some entities displayed in the original model, removed others, and also added new entities. The original model had 6 entities, while our model has 8 entities. The mentioned changes are shown in Table 2 below.

Table 2. Changes in entities for the new NER model.

Original Model	**Current Model**	**Situation**
PERSON	NAME	*Changed*
TIME	DATE	*Changed*
PLACE	ADDRESS	*Changed*
JURISPRUDENCE	-	*Removed*
LEGISLATION	-	*Removed*
ORGANIZATION	-	*Removed*
-	CPF	*Added*
-	PHONE	*Added*
-	EMAIL	*Added*
-	MONEY	*Added*
-	ZIP CODE	*Added*

For the new entities, we employed regular expressions (Regex) to enhance NER accuracy. Regex proved to be a valuable technique for identifying and replacing specific text patterns (highlighted in the study [5]), such as phone numbers, email addresses, and other personal identifiers, which are not always captured by standard NER methods. This complementary approach allowed for more granular filtering of added data, strengthening the effectiveness of identification when dealing with a wide range of formats.

5 Results

5.1 Evaluation of the New NER Model

Based on the generated dataset, we conducted evaluations using the F1-Score. The table below 3 contains the results obtained.

[12] https://huggingface.co/pierreguillou/ner-bert-large-cased-pt-lenerbr

Table 3. Evaluation Results

Entity	**Precision**	**Recall**	**F1-Score**	**Support**
NAME	0.73	0.83	0.777	948
DATE	0.913	0.994	0.952	856
ADDRESS	0.363	0.391	0.377	1135
CPF	0.988	1.0	0.994	887
PHONE	0.982	0.944	0.963	987
EMAIL	0.99	1.0	0.995	938
MONEY	0.971	1.0	0.985	965
ZIP CODE	1.0	0.526	0.69	851
Overall	0.837	0.826	0.832	-

We observed that the best-performing entity was "*EMAIL*" with an F1-Score of 0.995, while the worst was "*ADDRESS*" with an F1-Score of 0.377. It's important to note that the first is identified through Regex, while the latter by the NER model. With a simple analysis of the "*ADDRESS*", our assumption is that it contains text that could be confused with names of people, animals, among many others. There is also the fact that during the synthetic data generation process, the model may have generated addresses that are uncommon, which could have made correct identification more difficult.

The model achieved an overall F1-Score of 0.832, which is a satisfactory result for an NER model. The entity "*ZIP CODE*" had an F1-Score of 0.69, which is below expectations. This result can be attributed to the model's difficulty in correctly identifying ZIP codes, which consist of 8 digits and a hyphen. The hyphen was not captured by the initially created Regex pattern, as we had made the hyphen mandatory, hence a ZIP code like "*12345678*" was not recognized. To address this issue, we added a regular expression to also identify ZIP codes without a hyphen. The table 4 contains the results obtained after this correction.

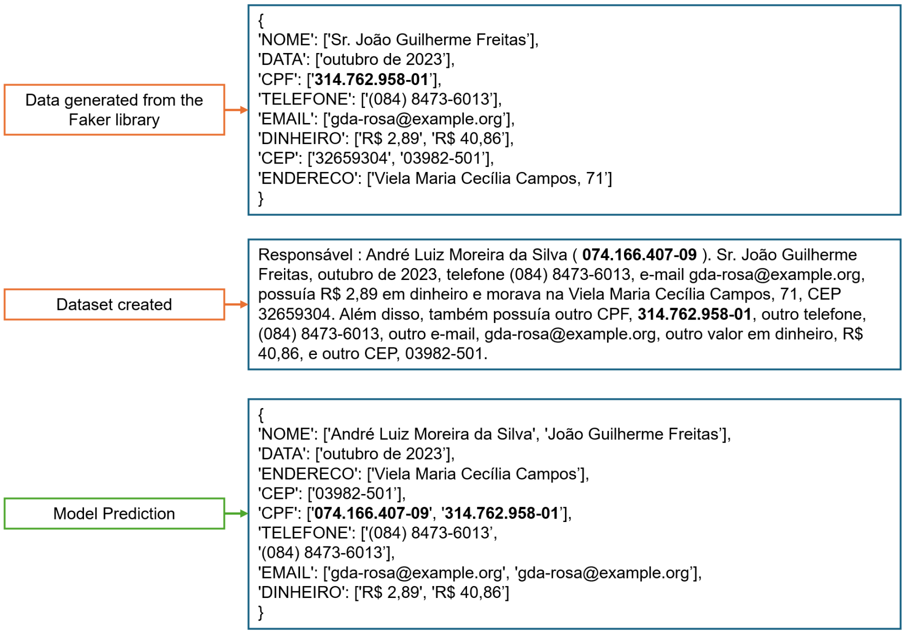
Table 4. Evaluation Results with Regex Corrections.

Entity	**Precision**	**Recall**	**F1-Score**	**Support**
NAME	0.73	0.83	0.777	948
DATE	0.913	0.994	0.952	856
ADDRESS	0.363	0.391	0.377	1135
CPF	0.988	1.0	0.994	887
PHONE	0.982	0.944	0.963	987
EMAIL	0.99	1.0	0.995	938
MONEY	0.971	1.0	0.985	965
ZIP CODE	1.0	0.999	0.999	851
Overall	0.845	0.879	0.862	-

By comparing the tables 3 and 4, we can see that the model correction with the addition of regular expressions for ZIP code identification significantly improved the F1-Score for this entity. The F1-Score for ZIP CODE increased from 0.69 to 0.99, which is an excellent result. The overall F1-Score of the model also improved, from 0.83 to 0.86. This demonstrates that adding regular expressions to correct entity identification errors can significantly enhance identification accuracy.

It is interesting to note that terms identified using Regex, such as "*CPF*", "*PHONE*", "*EMAIL*", "*MONEY*", and "*ZIP CODE*", achieved F1-Scores that were not always perfect. Upon further investigation, we observed some interesting behaviors. For this analysis, let's focus on the "*Entity*" column and the values identified with Regex.

Analyzing the "*Recall*" value for "*CPF*", we see a 100% score, meaning the model correctly identified all cases where "*CPF*" should not appear. However, for "*Precision*", the value was 98.2%, indicating the percentage correctly identified. Upon analyzing the cases where the model missed, we observed some instances like the one shown below in Figure 4.



```
{
'NOME': ['Sr. João Guilherme Freitas'],
'DATA': ['outubro de 2023'],
'CPF': ['314.762.958-01'],
'TELEFONE': ['(084) 8473-6013'],
'EMAIL': ['gda-rosa@example.org'],
'DINHEIRO': ['R$ 2,89', 'R$ 40,86'],
'CEP': ['32659304', '03982-501'],
'ENDERECO': ['Viela Maria Cecília Campos, 71']
}
```

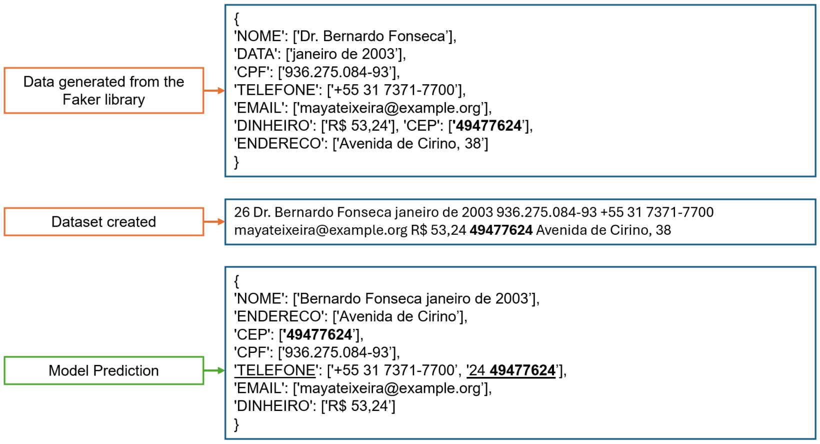
Responsável : André Luiz Moreira da Silva (**074.166.407-09**). Sr. João Guilherme Freitas, outubro de 2023, telefone (084) 8473-6013, e-mail gda-rosa@example.org, possuía R$ 2,89 em dinheiro e morava na Viela Maria Cecília Campos, 71, CEP 32659304. Além disso, também possuía outro CPF, **314.762.958-01**, outro telefone, (084) 8473-6013, outro e-mail, gda-rosa@example.org, outro valor em dinheiro, R$ 40,86, e outro CEP, 03982-501.

```
{
'NOME': ['André Luiz Moreira da Silva', 'João Guilherme Freitas'],
'DATA': ['outubro de 2023'],
'ENDERECO': ['Viela Maria Cecília Campos'],
'CEP': ['03982-501'],
'CPF': ['074.166.407-09', '314.762.958-01'],
'TELEFONE': ['(084) 8473-6013',
'(084) 8473-6013'],
'EMAIL': ['gda-rosa@example.org', 'gda-rosa@example.org'],
'DINHEIRO': ['R$ 2,89', 'R$ 40,86']
}
```

Fig. 4. Issues regarding the model's "*Precision*".

This indicates that, even though we had instructed the LLM to generate text without creating new terms, it did. This behavior repeats for the other Regex terms about "*Precision*". As for "*Recall*", the model correctly identified all cases, except for "*PHONE*", which scored 94.4%. In this case, our model made an error in just one instance, resulting in a false positive, as shown in the example in Figure 5.

It is still valid to say that our proposed model also ended up correctly identifying an additional data entity. In this way, we see a kind of inversion of our proposal, where our model ends up correcting the new dataset, improving its quality. This is a positive point, as it demonstrates that the model can be used to enhance data quality, even if that is not its primary objective.



```
{
'NOME': ['Dr. Bernardo Fonseca’],
'DATA': ['janeiro de 2003’],
'CPF': ['936.275.084-93’],
'TELEFONE': ['+55 31 7371-7700’],
'EMAIL': ['mayateixeira@example.org’],
'DINHEIRO': ['R$ 53,24'], 'CEP': ['49477624’],
'ENDERECO': ['Avenida de Cirino, 38’]
}
```

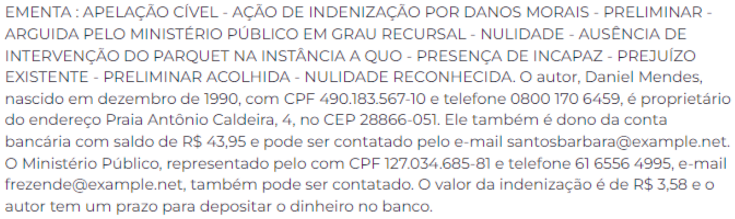
```
26 Dr. Bernardo Fonseca janeiro de 2003 936.275.084-93 +55 31 7371-7700
mayateixeira@example.org R$ 53,24 49477624 Avenida de Cirino, 38
```

```
{
'NOME': ['Bernardo Fonseca janeiro de 2003’],
'ENDERECO': ['Avenida de Cirino’],
'CEP': ['49477624’],
'CPF': ['936.275.084-93’],
'TELEFONE': ['+55 31 7371-7700’, '24 49477624’],
'EMAIL': ['mayateixeira@example.org’],
'DINHEIRO': ['R$ 53,24’]
}
```

Fig. 5. Issues regarding the model’s “*Recall*”

5.2 Usage Example of the New NER Model

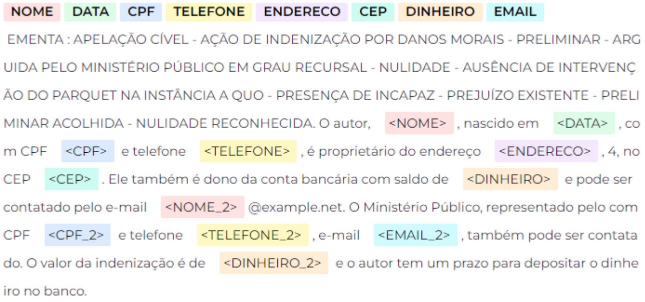
For the following example, we used the synthetic text generated as shown in Figure 3. We created an interface using Python to facilitate the pseudonymization of the text. Thus, we established one input field and two output fields. The input field is for the original text, and the output fields are for the identified terms and another for the pseudonymization secret. The Figure 6 below contains the original text, while Figures 7 and 8 contain data related to the generated outputs.



EMENTA : APELAÇÃO CÍVEL - AÇÃO DE INDENIZAÇÃO POR DANOS MORAIS - PRELIMINAR - ARGUIDA PELO MINISTÉRIO PÚBLICO EM GRAU RECURSAL - NULIDADE - AUSÊNCIA DE INTERVENÇÃO DO PARQUET NA INSTÂNCIA A QUO - PRESENÇA DE INCAPAZ - PREJUÍZO EXISTENTE - PRELIMINAR ACOLHIDA - NULIDADE RECONHECIDA. O autor, Daniel Mendes, nascido em dezembro de 1990, com CPF 490.183.567-10 e telefone 0800 170 6459, é proprietário do endereço Praia Antônio Caldeira, 4, no CEP 28866-051. Ele também é dono da conta bancária com saldo de R$ 43,95 e pode ser contatado pelo e-mail santosbarbara@example.net. O Ministério Público, representado pelo com CPF 127.034.685-81 e telefone 61 6556 4995, e-mail frezende@example.net, também pode ser contatado. O valor da indenização é de R$ 3,58 e o autor tem um prazo para depositar o dinheiro no banco.

Fig. 6. Input Data

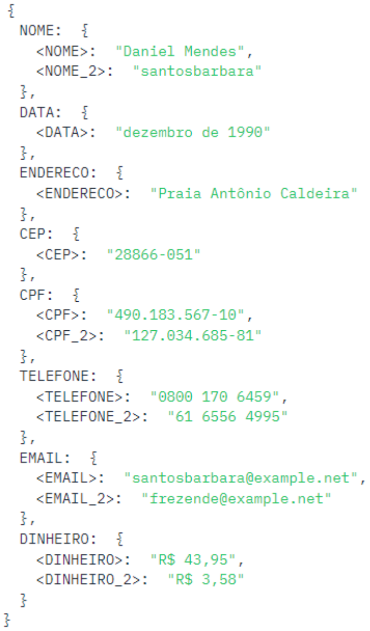
As shown in Figures 7 and 8 below, here again is the translation of the entities from Portuguese to English: "*NOME*" (NAME), "*DATA*" (DATE), "*ENDERECO*" (ADDRESS), "*TELEFONE*" (PHONE), "*DINHEIRO*" (MONEY), "*CEP*" (ZIP CODE).



NOME DATA CPF TELEFONE ENDERECO CEP DINHEIRO EMAIL

EMENTA : APELAÇÃO CÍVEL - AÇÃO DE INDENIZAÇÃO POR DANOS MORAIS - PRELIMINAR - ARGUIDA PELO MINISTÉRIO PÚBLICO EM GRAU RECURSAL - NULIDADE - AUSÊNCIA DE INTERVENÇÃO DO PARQUET NA INSTÂNCIA A QUO - PRESENÇA DE INCAPAZ - PREJUÍZO EXISTENTE - PRELIMINAR ACOLHIDA - NULIDADE RECONHECIDA. O autor, <NOME> , nascido em <DATA> , com CPF <CPF> e telefone <TELEFONE> , é proprietário do endereço <ENDERECO> , 4, no CEP <CEP> . Ele também é dono da conta bancária com saldo de <DINHEIRO> e pode ser contatado pelo e-mail <NOME_2> @example.net. O Ministério Público, representado pelo com CPF <CPF_2> e telefone <TELEFONE_2> , e-mail <EMAIL_2> , também pode ser contatado. O valor da indenização é de <DINHEIRO_2> e o autor tem um prazo para depositar o dinheiro no banco.

Fig. 7. Output Data - Entity Recognition



```
{
  NOME:  {
    <NOME>:  "Daniel Mendes",
    <NOME_2>:  "santosbarbara"
  },
  DATA:  {
    <DATA>:  "dezembro de 1990"
  },
  ENDERECO:  {
    <ENDERECO>:  "Praia Antônio Caldeira"
  },
  CEP:  {
    <CEP>:  "28866-051"
  },
  CPF:  {
    <CPF>:  "490.183.567-10",
    <CPF_2>:  "127.034.685-81"
  },
  TELEFONE:  {
    <TELEFONE>:  "0800 170 6459",
    <TELEFONE_2>:  "61 6556 4995"
  },
  EMAIL:  {
    <EMAIL>:  "santosbarbara@example.net",
    <EMAIL_2>:  "frezende@example.net"
  },
  DINHEIRO:  {
    <DINHEIRO>:  "R$ 43,95",
    <DINHEIRO_2>:  "R$ 3,58"
  }
}
```

Fig. 8. Output Data - Pseudonymization Secret

5.3 Limitations

As with any automated system, there are limitations that must be considered to ensure the effectiveness and accuracy of pseudonymization. One of these limitations is the need for human review. Our model may occasionally make errors in the pseudonymization of named entities. This requires a subsequent human review process to correct possible errors and ensure that the data is properly hidden, according to legal requirements.

Another significant limitation is the lack of cross-referencing between terms previously identified in the document. For example, if the system identifies "*Maria Silva*" as a person in one part of the text, it may not recognize "*Maria*" in a subsequent reference as the same person, potentially labeling it differently, such as "*NAME_2*". This problem, which seems simple to solve, may involve complex issues, such as homonyms and the overall context of the text.

Additionally, the model can generate false positives, which are errors where non-personal data are erroneously categorized with a different label. For example, a street name may be incorrectly identified as a person's name.

Finally, the model struggles to identify and correct typographical errors in critical information, such as CPF or phone numbers. Such typographical errors, which may include spaces or incorrect characters inserted between numbers, are not recognized by the system, potentially leading to a failure in properly concealing these sensitive data. This issue is likely to be addressed in subsequent work through optimization of the transformer model.

6 Conclusion

This study addressed the challenge of pseudonymizing Personally Identifiable Information (PII) in legal texts to comply with Brazil's General Data Protection Law (LGPD). We employed artificial intelligence technologies, including a Transformer model and Regex techniques, to effectively identify and conceal PII. Tests revealed that although our model showed good initial results, adjustments are necessary to enhance its precision and effectiveness.

The conclusions highlight that the combined approach of NER and Regex is promising, allowing for more accurate entity identification. However, methodological limitations include the need for ongoing adjustments to the model to handle atypical cases and the reliance on high-quality training data to maintain accuracy. Moreover, the technique faces challenges in consistently identifying all PII categories, with some entities like addresses showing lower results compared to others, such as CPFs and emails.

In summary, while initial results on synthetic data were promising, tests on real data are necessary to ensure the model's effectiveness in practical environments. The main advantage of the proposed model is its lightweight nature, allowing it to operate on systems with less computational capacity. This feature makes the model particularly valuable for applications in resource-limited environments or where processing speed is crucial.

7 Future Work

For future work, one of the main intentions is to mitigate the limitations of the current model, especially in terms of consistency in entity identification and error reduction, such as false positives and the non-cross-referencing of named entities. The idea is to refine the existing model and incorporate more precise feedback in the model training phase to improve its accuracy and efficiency.

Plans also include developing an interactive model that learns from corrections made by users. This active learning approach will allow the system to continuously refine its identification and categorization of PII, adjusting to the peculiarities of legal text and the specific pseudonymization preferences of the user.

Lastly, one of the focuses will be on identifying new terms related to PII that are continually emerging with changes in data collection practices and legislation. Recognizing that new types of personal data may arise, the model needs to be prepared and ready to respond to these new demands, enabling all forms of PII to be adequately identified and protected in accordance with the latest legal and ethical standards.

Acknowledgments. The authors acknowledge the use of the Llama version 3 model for generating data to enhance the dataset utilized in this research. The synthetic data generation included sensitive information to ensure the robustness of the model. The authors take full responsibility for the content of the article, including ensuring the originality and correctness of all text, and have taken all necessary precautions to ensure that no real sensitive information is used or exposed.

References

1. de Andrade, C.M., França, C., Belém, F., Jallais, G., Ganem, M.A., Texeira, G., Laender, A.H., Gonçalves, M.A.: Promptner: Uma abordagem para reconhecimento de entidades nomeadas em dados sensíveis a partir de instâncias rotuladas automaticamente. In: Anais do XXXVIII Simpósio Brasileiro de Bancos de Dados. pp. 269–281. SBC (2023)
2. Luz de Araujo, P.H., de Campos, T.E., de Oliveira, R.R., Stauffer, M., Couto, S., Bermejo, P.: Lener-br: a dataset for named entity recognition in brazilian legal text. In: Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13. pp. 313–323. Springer (2018)
3. de Dados Pessoais, P.: Lgpd (2020)
4. Fredriksson, T., Mattos, D.I., Bosch, J., Olsson, H.H.: Data labeling: An empirical investigation into industrial challenges and mitigation strategies. In: Product-Focused Software Process Improvement: 21st International Conference, PROFES 2020, Turin, Italy, November 25–27, 2020, Proceedings. p. 202–216. Springer-Verlag, Berlin, Heidelberg (2020). https://doi.org/10.1007/978-3-030-64148-1_13, https://doi.org/10.1007/978-3-030-64148-1_13
5. GR, G.M., Abhi, S., Agarwal, R.: A hybrid resume parser and matcher using regex and ner. In: 2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT). pp. 24–29. IEEE (2023)

6. Lewis, P.S.H., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks. CoRR **abs/2005.11401** (2020), `https://arxiv.org/abs/2005.11401`
7. Mota, C.C., Nascimento, A.C., Miranda, P.B., Mello, R.F., Maldonado, I.W., Coelho Filho, J.L.: Reconhecimento de entidades nomeadas em documentos jurídicos em português utilizando redes neurais. In: Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional. pp. 130–140. SBC (2021)
8. Nascimento, B.A.R., Botto Filho, M., Gomes, V.G., Menezes, H.K.A., Silva, A.N.M., et al.: Breve histório do cadastro de pessoa física–cpf e sua relação com a teoria dos números. Caderno de Graduação-Ciências Exatas e Tecnológicas-UNIT-SERGIPE **2**(3), 125–135 (2015)
9. Nunes, L.F.P., dos Santos, J.C.F.: Lgpd–uma visão de tecnologia e agnóstica. Revista Direito & Paz **2**(49), 218–237 (2023)
10. Oliveira, V., Nogueira, G., Faleiros, T., Marcacini, R.: Combining prompt-based language models and weak supervision for labeling named entity recognition on legal documents. Artificial Intelligence and Law pp. 1–21 (2024)
11. Patsakis, C., Lykousas, N.: Man vs the machine: The struggle for effective text anonymisation in the age of large language models. arXiv preprint arXiv:2303.12429 (2023)
12. Sakhaee, N., Wilson, M.C.: Information extraction framework to build legislation network. CoRR **abs/1812.01567** (2018), `http://arxiv.org/abs/1812.01567`
13. Zhang, S., He, L., Dragut, E., Vucetic, S.: How to invest my time: Lessons from human-in-the-loop entity extraction. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. p. 2305–2313. KDD '19, Association for Computing Machinery, New York, NY, USA (2019). `https://doi.org/10.1145/3292500.3330773`, `https://doi.org/10.1145/3292500.3330773`