

Name Pattern Recognition: A Model Proposal Applied to the Anonymization of Unstructured Data

Marisa de Andrade(✉), Kim Lima, Brenda Cardoso, João Lima, Renato Torres, and Nelson Neto

Federal University of Pará (UFPA), Belém, PA 66075-110, Brazil
{marisamoreno,renatohidaka,nelsonneto}@ufpa.br,
{kim.lima,brenda.cardoso,joao.costa.lima}@icen.ufpa.br

Abstract. The increasing digitalization of information has heightened concerns about the privacy of personal data, especially in unstructured textual documents. In this context, anonymization emerges as a crucial tool to ensure compliance with data protection regulations, such as the General Data Protection Law (LGPD) in Brazil. This paper presents a solution for the identification and anonymization of personal names in texts in Brazilian Portuguese, focusing on police reports (BOs), using natural language processing and machine learning techniques. Our approach employs a computational model that combines different machine learning techniques, integrating grammatical classification and context analysis to refine the classification of personal names and minimize the occurrence of false negatives. The results obtained were compared with BERTimbau and spaCy, reference models in the field. The methodology was evaluated using BOs provided by the Civil Police of Pará, achieving a recall of 99.62% in identifying personal names. While the increase in recall contributed to a rise in the number of false positives, this trade-off was necessary to maximize the coverage of personal names in the anonymization process. Furthermore, entropy analysis demonstrated that anonymization preserved textual variability, supporting the robustness of the proposal. The research results reinforce the methodology's potential to enhance privacy protection in unstructured texts, ensuring compliance with regulations and strengthening data security in sensitive contexts.

Keywords: Data Anonymization · Natural Language Processing · Machine Learning · Privacy Protection · Brazilian Portuguese

1 Introduction

The increasing generation of data, resulting from a significant digital transformation, has intensified the debate on rights and responsibilities in virtual environment [12], prompting governments to regulate the use of personal data [1]. In

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
O. Gervasi et al. (Eds.): ICCSA 2025, LNCS 15649, pp. 250–267, 2025.
https://doi.org/10.1007/978-3-031-96997-3_16

this context, several laws have been enacted to try to ensure a balance between technological advancement and the protection of individual privacy.

The European Union was the pioneer in implementing the General Data Protection Regulation (GDPR), serving as inspiration for similar legislation in various parts of the world, including Brazil's General Data Protection Law (LGPD) [17]. Like the GDPR, the LGPD[1] defines personal data as information that identifies or makes an individual identifiable. It established guidelines for the collection, storage, processing, and sharing of such data, ensuring transparency, security, and respect for privacy.

Anonymization, as defined in Article 5, Section XI of the LGPD, refers to the use of reasonable and available technical means at the time of processing, through which data loses the possibility of direct or indirect association with an individual. Various anonymization techniques are described in the literature, one of which attribute masking. This approach involves replacing characters of interest with predefined symbols, such as asterisks, providing the necessary level of anonymization without completely compromising the textual structure [14].

However, the additional challenge posed by unstructured data, such as news, social media posts, and police reports (BOs), lies not in anonymization itself but in correctly identifying the information that needs to be anonymized. Unlike structured data, where sensitive attributes are organized in predefined formats, free-text data exhibit linguistic variability, ambiguity, and context dependency, which significantly complicate the identification of personal information.

In the field of public security, the process of anonymizing personal data in BOs is a current issue. Automatically identifying personal names, for instance, can be a challenging task, as names may have identical spelling to common words, be subject to orthographic and regional variations, typing errors, as well as the issue of uncommon names [16]. The heterogeneity of unstructured texts in BOs and the ambiguity of natural language [18] significantly increase the complexity of the task. This ambiguity occurs when the same word takes on different meanings depending on the context, requiring advanced computational techniques for more robust identification.

In the context of the LGPD, failing to anonymize a personal name poses a critical risk, as it exposes sensitive information and compromises individuals' privacy. Minimizing the occurrence of false negatives is essential, as failing to anonymize a personal name could result in the improper disclosure of protected data. On the other hand, anonymizing a common word identified as a name does not constitute a data leak, making this type of error less harmful, emphasizing the importance of strategies that prioritize the complete recovery of personal names, ensuring a high recall in the anonymization process.

Given this scenario, this research proposes a solution to address ambiguities in the Portuguese language with the aim of anonymizing personal names in sensitive documents. For this purpose, a dataset provided by the Civil Police of the State of Pará (PC/PA) was used, containing 5,858 police reports. These documents are

[1] [Online]. Available: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. [Accessed: Sep. 13, 2024].

characterized by unstructured texts and contain detailed descriptions of criminal incidents, where the correct recognition of personal names is crucial for effective anonymization.

To address the challenges imposed by unstructured data, the proposed methodology employs natural language processing (NLP) algorithms and machine learning (ML) techniques. The solution includes two computational models. The first performs the initial identification of personal names, while the second improves the results of the first, reducing the occurrence of false negatives and, consequently, expanding the coverage of personal names. To achieve this, the proposed strategy combines contextual analysis with grammatical classification to mitigate the intrinsic ambiguities of Brazilian Portuguese (PB), thereby improving detection accuracy while minimizing information loss.

To evaluate the effectiveness of the solution, evaluation metrics such as accuracy, precision, recall, and F1-score were applied to a real dataset. Additionally, the BERTimbau and spaCy models, widely used in the literature, were also tested and compared, to evaluate the impact of our proposed approach on unstructured data identification. The results revealed that the proposed method achieved a recall of 99.62% in identifying personal names, demonstrating its effectiveness in mitigating false negatives. Furthermore, the entropy analysis of the original and anonymized texts showed a reduction of only 0.020 bits, indicating that the anonymization preserved the integrity of textual information.

These findings validate the potential use of intelligent algorithms for personal data protection, specifically in the process of anonymizing names in texts derived from police reports. Thus, anonymized texts retain their utility for future analyses, striking a balance between privacy protection and textual comprehensibility.

Given the regulatory requirements arising from the LGPD, this research represents a valuable contribution to the field of personal name anonymization, providing a scalable and efficient solution for handling sensitive data in unstructured texts. The solution reinforces the need for hybrid strategies that combine traditional techniques with machine learning-based approaches.

2 Related Work

In order to map research on personal data anonymization, searches were conducted in repositories such as Scopus, Elsevier, Scielo, SciTePress, and IEEE Xplore, as well as consultations with books and Brazilian regulations. The keywords used included data anonymization, LGPD, data protection in general, and personal data, considering only publications from 2020 onward that dealt with unstructured data.

In recent years, data anonymization has gained prominence across various domains, such as public safety and legal applications, driven by data protection regulations. In this context, the use of Named Entity Recognition (NER) has been explored in the legal field, as seen in [2,4]. In particular, [2] proposed a solution for the pseudonymization of legal data using NER, combining transformers and regular expressions to identify and mask personally identifiable

information (PII) in compliance with LGPD. In public safety, [9] applied ad hoc NER models embedded in the AGORA system, achieving an average F1-Score of 92.94% on reports from the Spanish National Police. The results highlight that customized models for specific contexts outperform generic solutions, although furhter research is needed to enhance the adaptability of these approaches.

NER solutions for targeting Portuguese-language files were proposed in [5]. The approach presented by the authors enables the identification and classification of named entities in historical documents. To achieve this, the work considered the use of CNNs, LSTMs, and Maximum Entropy models, resulting in the creation of a labeled dataset that is publicly available[2].

Regarding tasks involving NER in both structured and unstructured data, models such as BERT [6] and spaCy [8] have proven efficient, especially for identifying entities like names in texts. However, these approaches still face limitations when applied to unstructured data, where context plays a critical role [11,13].

In Brazil, data anonymization faces particular challenges such as regional variations, linguistic diversity, and the presence of names indigenous origin, requiring customized approaches to ensure the effectiveness of the process. To address these challenges, it is necessary to develop solutions tailored to the Brazilian context, respecting cultural and social characteristics, while adhering to the principles outlined in the LGPD [3].

In this regard, this research differs from other approaches, by not only recognizing named entities in unstructured data but also considering the context in which they appear. Our solution uses local corpus and integrates machine learning techniques and text preprocessing to anonymize personal data, tackling the ambiguity of natural language and regional name variation. The combination of these techniques with grammatical classification for context analysis allowed us to assess the effectiveness of the proposed methodology in real-world scenarios, demonstrating its feasibility and contribution to privacy protection in Brazil, considering its local specificities and data protection regulations.

3 Methodology

The proposed methodology for classifying personal names in unstructured texts was based on NLP and ML techniques, developed according to the flowchart in Fig. 1. More specifically, the methodology followed five main steps: (i) Collection and organization of the databases, including personal names, common words, and police reports; (ii) Selection of machine learning algorithms to classify personal names; (iii) Addition of contextual and grammatical features; (iv) Implementation of an adapted k-fold test with iterative sampling to handle data imbalance; (v) Training and evaluation of the proposed solution for the task of identifying personal names.

[2] [Online]. Available: https://huggingface.co/lfcc/bert-portuguese-ner. [Accessed: Feb. 28, 2025].

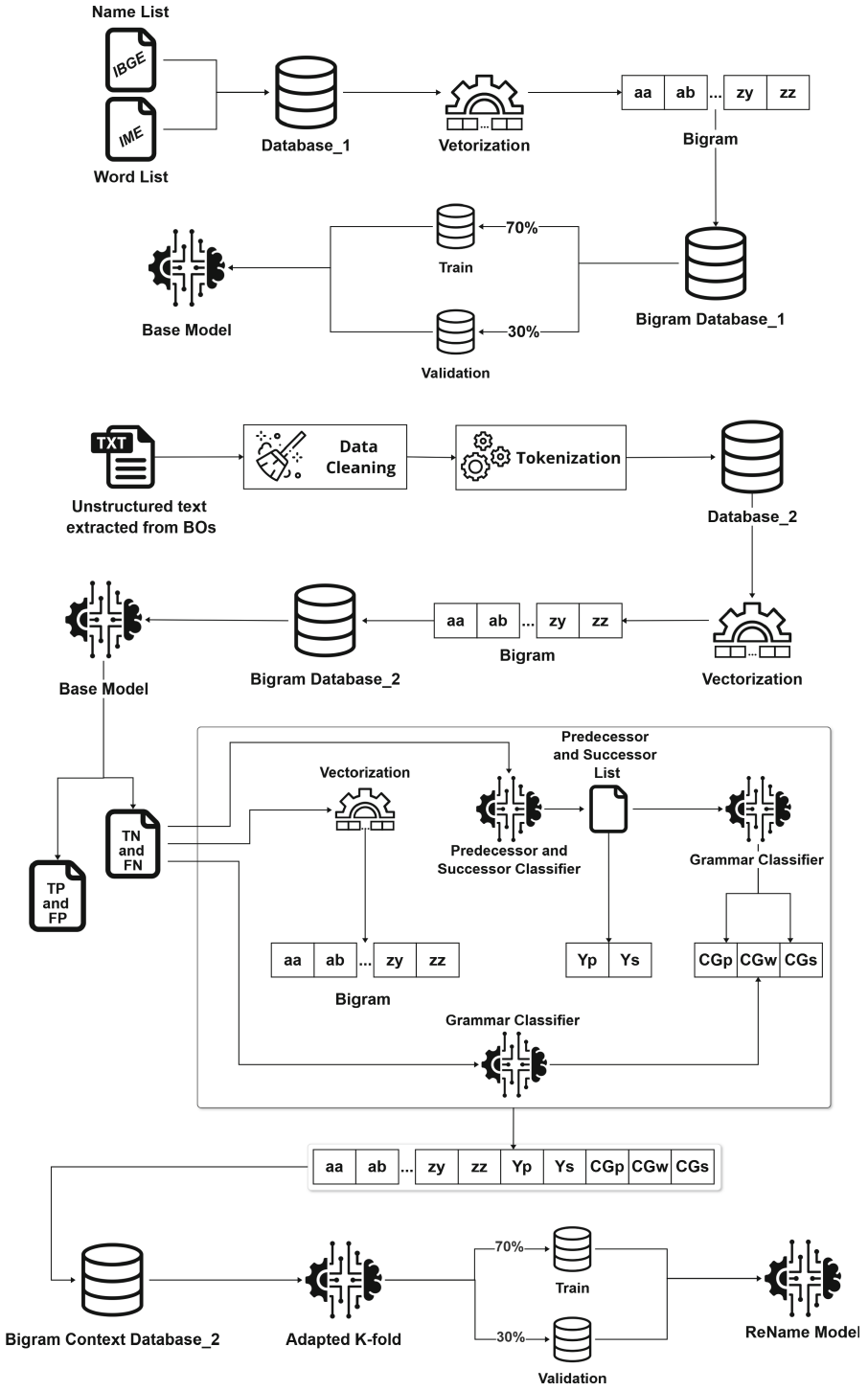


Fig. 1. Flowchart of the methodology used in the research.

3.1 Databases

Database_1. Initially, two lists were constructed. The first list is composed of personal names extracted from the Brasil.io platform[3], which provides data collected from the 2010 IBGE Demographic Census, and the second list is composed of common words in the Portuguese language provided by IME-USP[4]. The list of names contains 100,787 records from the "first_name" column, plus 1,774 records from the "alternative_names" column, totaling 102,561 names and their variations. For example, for the name "Davi", there are variations such as Davih, Davii, Davir, Davy, and Dhavi. The word list contains 245,019 terms, standardized without diacritical marks to maintain consistency with the name list.

After removing the intersection between the two lists, the Database_1 was assembled with 205,122 records, equally divided between common words (label 0) and personal names (label 1), as shown in Table 1. The words were randomly extracted from the IME-USP list, ensuring class balance and sample representativeness.

Table 1. Division of Database_1 into training and validation sets.

Dataset	Label 0 - Common Words	Label 1 - Personal Names	Total
Training	71,793	71,792	143,585
Validation	30,768	30,769	61,537
Total	102,561	102,561	205,122

Bigram Database_1. After splitting Database_1 into training (70%) and validation (30%), a vector was generated to represent each term in the dataset, facilitating its analysis in the modeling process [15]. In this work, the vectors were formed from bigrams of all possible combinations of the alphabet letters. Our hypothesis is that grapheme sequences in personal names and common words exhibit different patterns. Thus, the frequency of each bigram can be considered an indicator capable of distinguishing names from common words, making it a relevant feature in the classification task. This strategy captures more contextual information than unigrams, without the computational cost of trigrams and n-grams. For example, for the name "Davi", the bigrams DA, AV, and VI were set with the binary indicator 1, while all other bigrams are set to 0, as illustrated in Table 2. With this vectorization approach, the Bigram Database_1 was created with 205,122 rows and 677 columns, with 676 features and one column for the label.

[3] [Online]. Available: https://www.brasil.io/dataset/genero-nomes/. [Accessed: Aug. 20, 2024].

[4] [Online]. Available: https://www.ime.usp.br/~pf/dicios/. [Accessed: Aug. 20, 2024].

Table 2. Bigram feature vector for the term "Davi".

AA	AB	...	AV	...	DA	...	VI	...	ZY	ZZ
0	0	0	1	0	1	0	1	0	0	0

Database_2. This dataset was built from two batches of police reports (BOs) filed by the PC/PA in 2021, collected from several municipalities in the State of Pará, including Belém, Altamira, Marituba, Ananindeua, and Castanhal. The first batch includes 300 BOs previously labeled by the PC/PA, while the second batch contains another 471 BOs manually labeled by academic researchers. The composition of Database_2 in common words and personal names is shown in Table 3. During the preprocessing, stopwords were removed using the NLTK tool[5] to reduce the vocabulary size and eliminate irrelevant words, such as articles, prepositions, and conjunctions. This procedure was also applied to surnames, such as in "João da Silva" where anonymizing the word "da" does not cause issues. After this cleaning process, the data was tokenized.

Table 3. Composition of Database_2 by BO batches.

BO Source	Number of BOs	Common Words	Personal Names	Total
First batch	300	4,853	743	5,596
Second batch	471	48,156	4,526	52,682
Total	771	53,009	5,269	58,278

Bigram Database_2. The Database_2 then underwent a vectorization process similar to that applied to Database_1. As a result, Bigram Database_2 was generated.

3.2 Machine Learning Algorithm Selection

All algorithms were imported from the Scikit-learn library[6], version 1.3.2, using default parameters. This approach was adopted because the focus of this work is not on defining the optimal algorithm but rather to evaluate the performance of machine learning when applied to the data anonymization process, specifically in the task of classifying personal names.

Thus, four computational models, each based on a different algorithm, were trained using 70% of Bigram Database_1. Subsequently, two validation experiments were conducted. The first used the remaining 30% of Bigram Database_1

[5] [Online]. Available: https://www.nltk.org. [Accessed: Oct. 18, 2024].
[6] [Online]. Available: https://www.scikit-learn.org. [Accessed: Oct. 18, 2024].

(Experiment 1), while the second employed the 300 BOs comprising the first batch of Bigram Database_2 (Experiment 2). The goal was to investigate the ability of the computational models to identify automatically whether a sequence of characters corresponds to a personal name or a common word in different environments. The results obtained from Experiment 1 are presented in Table 4 and the confusion matrix in Fig. 2.

Table 4. Metrics calculated for each model in Experiment 1.

Model	Accuracy	Precision	Recall	F1-Score
SVM	0.931180	0.911709	0.958523	0.932768
Random Forest	**0.967970**	**0.961742**	**0.974714**	**0.968185**
Logistic Regression	0.940442	0.933080	0.948940	0.940943
Gradient Boosting	0.901896	0.875611	0.936882	0.905211

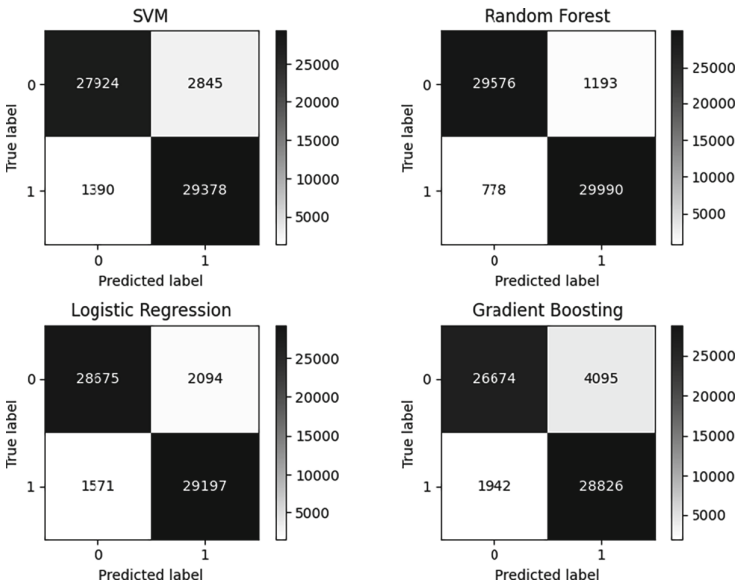


Fig. 2. Confusion Matrix for Experiment 1: Label 1 identifies personal names, and label 0 identifies common words in Brazilian Portuguese.

As show in Table 4, the Random Forest algorithm stood out, achieving over 96% across all evaluation metrics. Further analysis of the confusion matrix in Fig. 2 reveals that Random Forest correctly classified 29,990 instances from the validation set were correctly classified as personal names. In contrast, the SVM showed 2,845 false positive cases (instances classified as positive, but actually

negative) and 1,390 false negative cases (instances classified as negative, but actually positive).

An important point that should be noted is that the models proved effective in minimizing false negatives, that is, classifying personal names as common words. This is a crucial aspect of the problem motivating this work. This quality provides more confidence to the operator, as the likelihood of not anonymizing a personal name is reduced, thus increasing the reliability of the system. The analysis of false negatives in Random Forest, which stood out compared to the others, revealed typical issues from the literature, such as ambiguity, typographical errors, and uncommon or foreign names. Some examples are presented in Table 5.

Table 5. Examples of personal names that were incorrectly classified as common words by the Random Forest model in Experiment 1.

Possible Causes	Examples
Ambiguity	SOCORRO, TECLA, CORADO, LIVRAMENTO, AGRARIO, COLETA, ALEGAR, ADRENALINA, ALEGRIA, FALCAO, TRANQUILO, INFANCIA
Typing errors	FRABCISCA, FARNCISCO, ESTLA, MRCOS
Uncommon Names	CROMACIO, EXUPERIO, GONTRAN, TUDE
Indigenous-origin Names	ARAPUA, APOEMA, TUPA, POTIRA, POTY, IRAE
Foreign Names	FARUK, FAROUK, HASSEN, KARAM, ANDREAS

Most false negatives appear to stem from lexical ambiguity, where a word's meaning varies depending on context. For instance, the name "ALEGRIA" may have been misclassified due to its association with the emotion of the same name. The second most common cause includes typographical errors, such as "MRCOS", which was likely intended to be "MARCOS". Although less frequent, foreign, atypical, and indigenous names were also observed. Regarding the classification of indigenous names, even though it failed with "POTIRA" and "APOEMA", the model correctly identified the names "JACY", "ITAJACI" and "ARACI".

Continuing with the validation process, a second experiment was conducted in a more realistic and less controlled environment. Thus, Experiment 2 studied the performance of the Random Forest model (now referred to as the Base Model because it stood out in Experiment 1) when applied to texts from 300 police report statements (see Table 3). Ultimately, the Base Model correctly classified 674 out of the 743 instances of personal names, as shown in the confusion matrix in Fig. 3.

However, the evaluation metrics declined significantly compared to Experiment 1, particularly in terms of precision and F1-score, as shown in Table 6. This drop in performance is likely due to the high number of false positives (=814) resulting from the imbalance in the validation set, which contains about six times more words than names. Nonetheless, the accuracy and recall results remained promising.

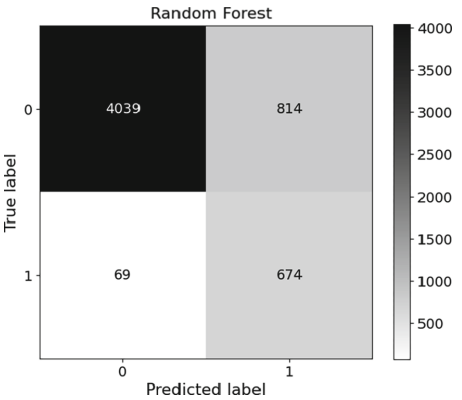


Fig. 3. Confusion matrix of Experiment 2: Label 1 identifies personal names, and label 0 identifies common words in Brazilian Portuguese. In this second experiment, only the Random Forest model was validated because it was highlighted in Experiment 1. For this reason, this model was developed as a basis for the following analyses.

Table 6. Metrics calculated for the Base Model in Experiment 2.

Model	Accuracy	Precision	Recall	F1-Score
Base	0.855969	0.476787	0.870794	0.616190

Regarding false negatives, lexical ambiguity is the primary cause, as shown in Table 7. For example, personal names such as "COELHO" and "CARNEIRO" were possibly associated with the animals of the same name. Typing errors, such as "PIMHO" instead of "PINHO", also influenced the classification errors. Furthermore, uncommon and foreign names continued to pose challenges for classification, such as "VASMORBITA" and "PEDTROVICH". In contrast, names of indigenous origin were correctly classified, demonstrating the good performance of the solution in these specific cases.

Table 7. Examples of personal names that were incorrectly classified as common words by the Base Model in Experiment 2.

Possible causes	Examples
Ambiguity	GRATIDAO, TORRES, BARRA, SANTOS, PRETO, FILHO, COELHO, CARNEIRO, COLARES, GUERREIRO, CRUZ, PIO
Typing errors	GONCACVES, PIMHO
Uncommon Names	VASMORBITA, RAMPIN, ROBISPIERRE
Foreign Names	PEDTROVICH, KYRIACOPOULOS, BASEGGIO, CUTRIM

3.3 Addition of Contextual and Grammatical Features

To initiate the feature generation process with contextual and grammatical information, Bigram Database_2 was submitted, as an instance, to the Base Model. The list of words resulting from the prediction was then divided into two groups: words that were correctly predicted as words (=33,746 TN); and names that were incorrectly predicted as words (=269 FN).

After separating the predicted instances into TN and FN lists, a bigram vector was constructed for each term, following the same structure as the Database_1. In addition to the bigrams, other information, such as predecessor, successor, and grammatical classification was incorporated to capture the linguistic context.

As a result of this process, the feature vector was structured into three parts: (i) the bigram vector of the term labeled as FN or TN; (ii) the prediction of the predecessor (Y_p) and the successor (Y_s) of the term in the police report text; and (iii) the grammatical classification of the predecessor (CG_p), the term itself (CG_w), and the successor (CG_s), as exemplified in Table 8. This new database was named Bigram Context Database_2.

Table 8. Example of vectorization for the instance "BARRETO", predicted as a word (0). The predecessor and successor are nouns (1); if it were a word, it would receive label 0. The numbers 12 and 13 represent the grammatical classes noun and verb, respectively.

AA	AB	...	AR	...	ZY	ZZ	Y_p	Y_s	CG_p	CG_w	CG_s
0	0	0	1	0	0	0	1	1	12	12	13

The grammatical classification of the terms was performed using the NLP model from spaCy[7], considering 14 categories, including: noun, verb, and adjective. In cases where a predecessor or successor was absent, the label NULL was assigned to maintain data integrity and ensure a consistent feature structure.

3.4 Adapted K-Fold Test

During the creation of Bigram Context Database_2, an an imbalance between the VN and FN classes was identified, and training a computational model with this unbalanced distribution could lead to biased results, with the word class dominating the predictions. To mitigate this issue, the k-fold test was combined with an iterative sampling strategy to assess the variance resulting from the selection of random subgroups. Thus, in each of the 125 iterations, 269 samples from the VN class were randomly extracted, which is equivalent to the number of FN. The process was halted when the remaining number of VN samples fell below this threshold, leading to the exclusion of the last 121 samples from the

[7] [Online]. Available: https://www.spacy.io/api/doc. [Accessed: Nov. 05, 2024].

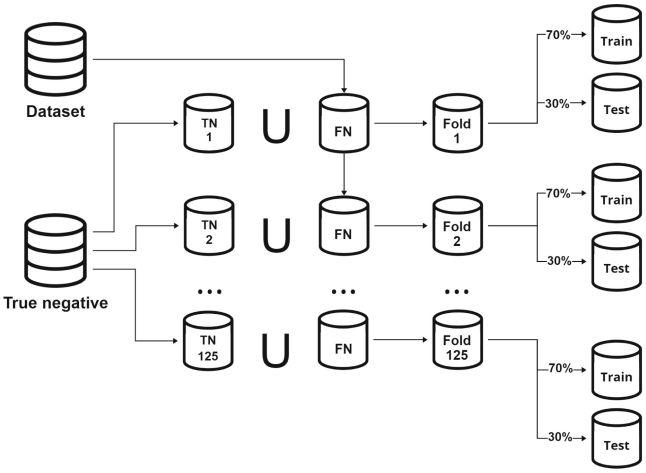


Fig. 4. Flow of the k-fold test applied to the dataset.

dataset. The workflow of the k-fold test performed with the Random Forest algorithm is shown in Fig. 4.
The analysis of the k-fold test results indicated relatively low variance, suggesting that the model demonstrated good generalization capability, even when exposed to randomly selected subgroups, as shown in Fig. 5.

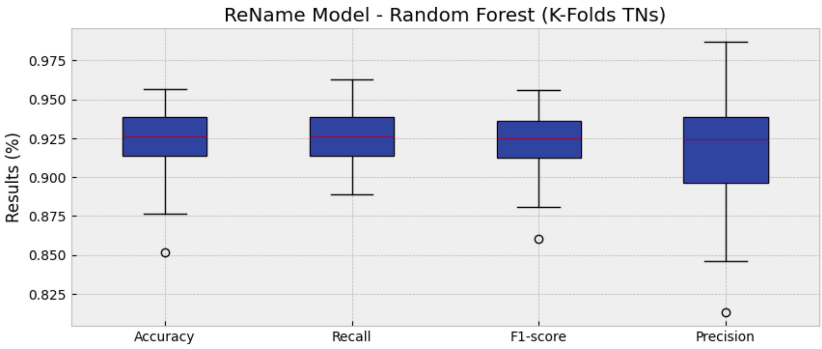


Fig. 5. Variance obtained for true negatives with the k-fold test.

3.5 Training and Validation of the ReName Model

To enhance the performance of the Base Model, with a particular focus on reducing false negatives, another computational model, called the ReName Model, was incorporated into the proposed solution.

The random sampling strategy was maintained, and 538 records (269 FN and 269 VN) from the Bigram Context Database_2 were used for the training and validation of the ReName Model. The training phase employed 70% of the dataset using the Random Forest algorithm, while the remaining 30% was reserved for validation. In this phase, the model correctly identified 78 unique instances as personal names, with only three misclassifications, as shown in the confusion matrix in Fig. 6. Other results can be found in Table 9. All metrics exceeded 90%, with a highlight for a recall of 96.29%.

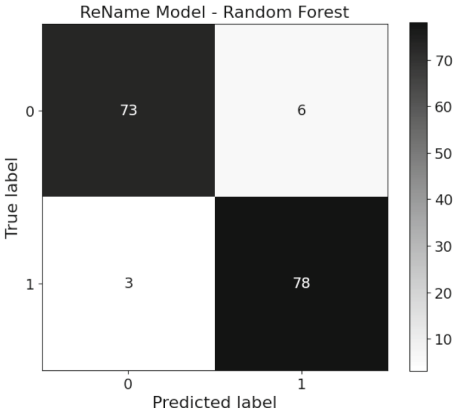


Fig. 6. Confusion matrix for the validation of the ReName Model: Label 1 identifies personal names, and label 0 identifies common words in Brazilian Portuguese.

Table 9. Metrics calculated in the validation of the ReName Model.

Model	Accuracy	Precision	Recall	F1-Score
ReName	0.94375	0.928571	0.962963	0.945455

The LIME (Local Interpretable Model-agnostic Explanations) explainability framework, widely used to interpret black-box machine learning algorithms [10], was employed to analyze the decisions made by the ReName Model. Through a Python API[8], the most influential features in specific classification decisions were identified.

To support this interpretation, several instances from the validation dataset were analyzed, considering the 10 most relevant features. In the case of the instance "TORRES", for example, correctly classified as a personal name, LIME assigned a 97% probability to class 1 (personal names) and 3% to class 0 (common words), as illustrated in Fig. 7.

[8] [Online]. Available: https://lime-ml.readthedocs.io. [Accessed: Nov. 20, 2024].

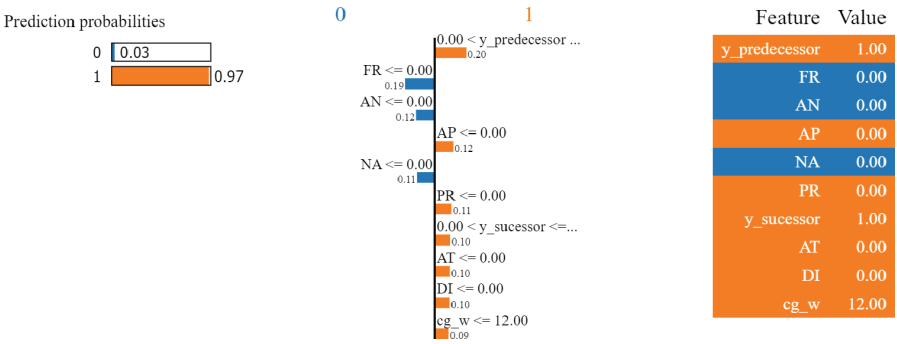


Fig. 7. Example of an instance correctly classified as a personal name based on the explainability provided by LIME.

It is important to note that the instance "TORRES" was incorrectly classified by the Base Model (see Table 7). This result reflects the influence of three features: predecessor (Y_p), successor (Y_s), and grammatical classification of the instance (CG_w). Both Y_p and Y_s indicated that the neighboring instances of "TORRES" were also personal names, contributing to the identification of sequential patterns. Meanwhile, CG_w received the number 12 (noun), reinforcing the correct prediction.

Next, the results of an evaluation conducted with our proposal are presented, analyzing its performance in classifying names and words in police reports. Additionally, the results are compared with two references from the literature: BERT and spaCy, enabling a more comprehensive analysis of the performance and advantages of the developed solution.

4 Evaluation

The evaluation of the proposed solution was conducted to assess its performance in classifying a dataset that was not seen during the training and validation phases. For this purpose, 231 preprocessed BO reports manually labeled by academic researchers, containing 2,135 personal names and 21,604 common words, were used as the test set. For comparative analysis, this same set was evaluated using the spaCy and BERTimbau models. The classification metrics obtained can be seen in Table 10.

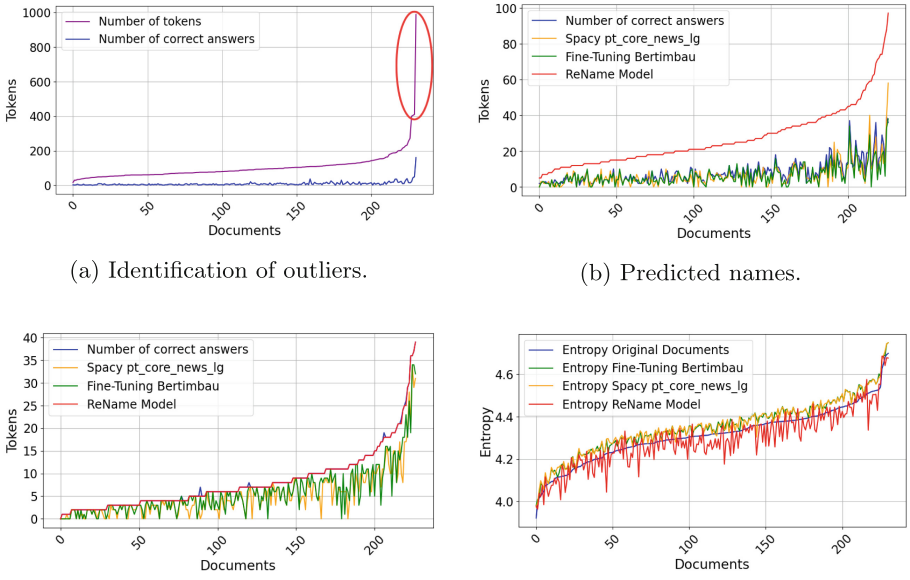
The correct classification of personal names in unstructured documents is crucial for ensuring compliance with the LGPD, as failure to anonymize a personal name may compromise individuals' privacy by improperly exposing sensitive information. Thus, the proposed solution was designed to minimize the number of non-anonymized names, even if this objective results in a higher number of common words being mistakenly classified as names.

The results of this evaluation are organized in Fig. 8. Initially, four documents with more than 400 tokens were identified as outliers (see Fig. 8a) and removed

Table 10. Performance metrics of the proposed solution, BERTimbau and spaCy.

Model	Accuracy	Recall	Precision	F1-Score
Our proposal	0.799780	0.996252	0.309516	0.472299
BERTimbau	0.963730	0.681498	0.889364	0.771678
spaCy	0.948312	0.612177	0.766119	0.680551

from the analysis of predictions and correct classifications to enhance result visualization, totaling 227 documents in the charts of Figs. 8b and 8c, respectively. However, for the entropy calculation of the predictions, presented in Fig. 8d, all 231 documents were considered to ensure a comprehensive analysis of the test set.



(a) Identification of outliers.

(b) Predicted names.

(c) Correctly classified cames.

(d) Entropy of Predictions.

Fig. 8. Distribution of tokens per document, ordered from the smallest to the largest. The number of correct responses corresponds to the number of tokens manually labeled as personal names in each document. (a) Identification of outliers in the number of tokens. (b) Number of tokens predicted as names. (c) Number of tokens correctly classified as names. (d) Entropy of the predictions, comparing the original and anonymized documents.

The number of tokens predicted as personal names by each model is shown in Fig. 8b. It is observed that the proposed solution presented the highest number of predictions in all BOs, indicating its high sensitivity to categorizing personal

names, as evidenced by its high recall (= 99.62%). However, the BERTimbau and spaCy models provided a more precise approach.

This difference in the behavior of the model is reflected in the values shown in Fig. 8c, which presents the number of tokens correctly classified as personal names. While our proposed solution delivered the expected performance, the BERTimbau and spaCy models compromised personal data protection by failing to identify a greater number of names.

It is essential to emphasize that maximizing the identification of personal names must not hinder the understanding of the report's context, which may occur if a sufficiently large number of common words are mistakenly anonymized. As presented in Fig. 8d, the entropy of the predictions [7] was calculated to provide a measure of uncertainty associated with the classifications and to allow a comparison between the original and anonymized documents. The reduction of 0.020 bits in the average entropy considering the proposed solution (from 4.315 to 4.295) demonstrates a low impact on the complexity of the anonymized texts. In contrast, BERTimbau and spaCy exhibited the opposite behavior, increasing the entropy of anonymized reports by 0.038 bits and 0.039 bits, respectively.

Experiment 1 can be replicated using the code made available on the Github platform[9]. Unfortunately, due to confidentiality issues with the BOs, this could not be done for the other experiments. All tests were conducted on a notebook with an AMD Ryzen 5500U processor (6 cores and 12 threads) and 12 GB of RAM, without a dedicated GPU.

5 Conclusion

This work presents a solution for identifying and anonymizing personal names in unstructured texts, with a focus on police reports written in Brazilian Portuguese. The solution combines machine learning and natural language processing techniques, structured in five stages, including data collection and organization, algorithm selection for building computational models, addition of contextual and linguistic features, and finally, training and evaluation of the solution. Among these stages, the construction of the Base Model, responsible for the initial identification of personal names, and the ReName Model, which improves classification by reducing false negatives, stand out.

Our approach was evaluated using police reports provided by PC/PA and achieved a recall of 99.62%, ensuring broad coverage of personal names in the texts. However, this result was achieved at the cost of a higher occurrence of false positives, which could compromise accuracy in other scenarios.

For comparison purposes, the widely recognized BERTimbau and spaCy models were also evaluated. In this regard, the results demonstrated that while our solution achieved the highest recall, the reference models showed a better balance between precision and recall, resulting in lower false positive rates.

[9] [Online]. Github repository provided by the authors.
Available: https://github.com/morenomarisa/namesClassification.

Thus, the choice between the models should consider the application's goal. When the priority is to ensure the anonymization of personal names, the ReName Model is best suits, as it maximizes the identification and protection of these entities. However, for the general purpose of name classification in unstructured data, the BERTimbau and spaCy models are more suitable, as they offer higher accuracy and reduce the misclassification of common words.

Additionally, an entropy analysis indicated minimal impact between the original and anonymized reports, preserving the textual structure of these documents. On the other hand, the texts anonymized by BERTimbau and spaCy showed a greater impact compared to the original texts. This finding confirms the importance of solutions that maintain textual consistency without compromising privacy protection, as in the proposed solution.

Although our approach prioritizes the maximum identification and anonymization of personal names, with a reduction in false negatives, the results show that there is room for optimizing the balance between recall and precision. As a future development, deep learning techniques could be adopted to enhance model accuracy and improve the balance between recall and precision results.

Ultimately, the solution presented in this work proves to be promising for applications requiring secure and efficient anonymization of textual documents, such as in public security and contexts requiring compliance with the LGPD.

References

1. Almeida, S.d.C.D.d., Soares, T.A.: Os impactos da lei geral de proteção de dados-lgpd no cenário digital. Perspectivas em Ciência da Informação **27**(03), 26–45 (2022)
2. Anselmo, M., Ribas, B.C.: Pseudonymization in legal texts according to the LGPD: A named entity recognition approach. In: Brazilian Conference on Intelligent Systems, pp. 309–323. Springer (2024)
3. Carvalho, A.P., Canedo, E.D., Carvalho, F.P., Carvalho, P.H.P.: Anonymisation and compliance to protection data: Impacts and challenges into big data. In: ICEIS (1), pp. 31–41 (2020)
4. Csányi, G.M., Nagy, D., Vági, R., Vadász, J.P., Orosz, T.: Challenges and open problems of legal document anonymization. Symmetry **13**(8), 1490 (2021)
5. Cunha, L.F.d.C., Ramalho, J.C.: NER in archival finding aids: extended. Mach. Learn. Knowl. Extract. **4**(1), 42–65 (2022)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 4171–4186 (2019)
7. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. In: Advances in Neural Information Processing Systems, vol. 17 (2004)
8. Honnibal, M., Montani, I.: spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear **7**(1), 411–420 (2017)
9. Juez-Hernandez, R., Quijano-Sánchez, L., Liberatore, F., Gómez, J.: Agora: an intelligent system for the anonymization, information extraction and automatic mapping of sensitive documents. Appl. Soft Comput. **145**, 110540 (2023)

10. Laguna, S., et al.: Explimeable: a visual analytics approach for exploring lime. In: 2023 Workshop on Visual Analytics in Healthcare (VAHC), pp. 27–33. IEEE (2023)
11. Li, X., Feng, J., Meng, Y., Han, Q., Wu, F., Li, J.: A unified MRC framework for named entity recognition. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5849–5859 (2020)
12. Lorenzon, L.N.: Análise comparada entre regulamentações de dados pessoais no brasil e na união europeia (lgpd e gdpr) e seus respectivos instrumentos de enforcement. Revista do Programa de Direito da União Europeia **1**, 39–52 (2021)
13. Luoma, J., Pyysalo, S.: Exploring cross-sentence contexts for named entity recognition with Bert. In: Proceedings of the 28th International Conference on Computational Linguistics pp. 904–914 (2020)
14. Nelson, G.S.: Practical implications of sharing data: a primer on data privacy, anonymization, and de-identification. In: SAS global forum proceedings, pp. 1–23 (2015)
15. Silva, M.P.d., Viera, A.F.G.: Descoberta de conhecimento com uso de técnicas de mineração de textos aplicadas em documentos textuais da investigação policial brasileira. Investigación bibliotecológica **35**(88), 161–183 (2021)
16. Silveira, M., et al.: An anonymization service for privacy in data mining. In: Proceedings of the 12th Latin-American Symposium on Dependable and Secure Computing pp. 214–219 (2023)
17. Sousa, T.R., Coutinho, M., Coutinho, L., Albuquerque, R.: Lgpd: Levantamento de técnicas criptográficas e de anonimização para proteção de bases de dados. In: Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais (SBSeg), pp. 55–68. SBC (2020)
18. Weitzenboeck, E.M., Lison, P., Cyndecka, M., Langford, M.: The GDPR and unstructured data: is anonymization possible? Int. Data Privacy Law **12**(3), 184–206 (2022)