DOI: 10.5748/19CONTECSI/PSE/DSC/7142

A SUPERVISED CLASSIFIER FOR POLICE REPORTS AT THE STATE OF PARÁ, BRAZIL

Helder Mateus dos Reis Matos ; https://orcid.org/0000-0002-5632-7948
Universidade Federal do Pará

Samara Lima de Souza ; https://orcid.org/0000-0001-9807-2595
Universidade Federal do Pará

Reginaldo Cordeiro dos Santos Filho ; https://orcid.org/0000-0002-0456-8547
Universidade Federal do Pará

João Crisóstomo Weyl Albuquerque Costa ; https://orcid.org/0000-0003-4482-6886
Universidade Federal do Pará

A SUPERVISED CLASSIFIER FOR POLICE REPORTS AT THE STATE OF PARÁ, BRAZIL

ABSTRACT

This paper describes the development of a supervised classifier constructed upon knowledge extracted from police report public databases, in the years between 2019 and 2021 in the state of Pará, Brazil. The classifier achieved an accuracy of approximately 78% for the prediction of 463 unique labels related to public safety. The resulting model can be used to improve the statistical processes of criminal analysts, both in quantitative and qualitative terms.

Keywords:
Machine Learning, Data Mining, Public Safety, Natural Language Processing, Text Classification.

1. Introduction

Public safety is one of the sectors of higher interest for the public administration of modern societies. The effort in maintaining the public order can be tracked back in the philosophy of legal sciences, described as the study of an organic and systematic set of rules which includes public and criminal law (Souza D. C., 1972).

In Brazil, the National System of Public Safety Information (SINESP), managed by the Ministry of Public Security, is the main infrastructure of data and criminal information in the country, handling data of penitentiary systems, firearms and ammunition tracking, genetic material, fingerprints, and drugs (Brasil, 2022). Created in 2012, SINESP is a secure and standardized informational system, that facilitates communication between members of the Unified Public Security System (SUSP). Each state of the Brazilian federation is responsible for the collection, normalization, and delivery of data extracted from police stations throughout the country. This integration requires a higher amount of effort, especially in the countryside where such systems tend to fail due to the lack of maintenance, computerized police stations, and capable professionals able to register and consolidate a massive volume of police records.

The growth in the amount of data generated in our society became the catalyst for the digitization of processes across the different public sectors. Most recently, the application of data science and data mining tools became a trend to be observed by those sectors. Public safety has a huge potential to acquire benefits generated by automatic tools of knowledge extraction in databases, including classification of police reports, the discovery of associative rules out of past events, information visualization, business intelligence, and extraction and generation of hidden attributes in public databases.

Under those circumstances, the current paper aims to describe the development of a supervised classifier of police reports, using data gathered from the Assistant Secretariat for Intelligence and Criminal Analysis (SIAC), an institution linked to the Public Security and Social Defense Secretariat (SEGUP) of the state of Pará, Brazil. An investigation of the tools used in data mining applied to textual data was conducted, which aided in the construction of a tool to process police station reports to predict a type of event, with the premise of being generalizable to different scenarios and jurisprudence. As for the resulting classifier model, it has the potential to be a mechanism to accelerate the statistical processes of the secretariat, along with an automatized qualitative analysis of a huge amount of criminal data.

This paper is organized as follows: Section 2 investigates related work to the scope of the current research. Section 3 describes the proposed methodology. Section 4 presents the obtained results and its general discussion. The paper is concluded in section 0, along with potential future work to be explored.

2. Related Work

This section summarizes literature related to the scope of supervised machine learning applied to textual data, which aided to determine which concepts and tools are used in the field of research.

A data science-driven approach focused on human interaction was proposed by (Qazi & Wong, 2019), being capable of extracting associations between crime patterns, organizing clusters of similar crimes, and identifying perpetrator networks and suspect lists based on spatial-temporal similarities and *modus operandi*. A robbery data set, with approximately 1.6 million records included personal information of perpetrators and victims, thus this information was submitted to an encryption step, making it impossible to link a person with a data sample. The proposed analysis generated a crime pattern visualization tool in a bi-dimensional space with dynamic attribute selection. The research also pointed out that patterns related to “uPVC Door” or “windows” were likely associated with a robbery.

An exploratory data analysis over police records limited to mobile phone robbery data in São Paulo between 2010 and 2018 found patterns using statistical, along with inferential, predictive, and spatial experiments (Vargas, 2019). The results pointed out that in nocturnal events the perpetrator is likely to use a motorbike, a pattern found in 84% of the events. Among predictive algorithms such as Logistic Regression, Decision Tree, Support-Vector Machine, Neural Networks, Random Forest, and Bagging, the last two were able to predict around 60,5% of the offenses.

Situated in Belém, Brazil, the authors in (Trindade, 2019) verified the relationship between murder and social vulnerability factors for the age group between 15 and 19 years old, from the years of 2008 until 2019. A quantitative analysis based on statistical techniques, able to expose interesting indexes and variables such as Correspondence Analysis and Principal Component Analysis, coupled with a qualitative analysis, able to interpret concepts and phenomena existent in the collected data, pointed out an absence of efforts of integration between healthcare, education, and public security, in addition to the highlighting of causes and risk factors to the occurrence of murder, directly affected by public policies, notably education, healthcare, housing, and employment. Another key point was noticed for the Metropolitan Region of Belém, where the profile of young individuals, male, single, lower education profile, students, freelancers, unoccupied, or part of the informal market was prone to be affected by murdering crimes.

The application of data science to the public security of the state of Rio de Janeiro was the focus of (Souza J. R., 2018), where data on the monthly evolution of records for the civil police stations were analyzed for the years between 2003 and 2018. The applied methodology consisted of data integration, data cleaning, and feature engineering steps, in addition to an exploratory data analysis that described indexes for criminal offenses throughout the mentioned period - in particular for the behavior of the current state government administration. Moreover, visualization tools for geographical crime distribution and the total amount of crimes correlated with demographic data were constructed.

The criminality in the state of Pará, Brazil, was studied in (Regateiro, Ramos, Souza, & Mello, 2021), for the years between 2017 to 2019. Records for crimes such as theft, robbery, vehicle robbery, murder, robbery with homicide, and bodily lesion followed by death, were collected and submitted to an exploratory analysis that generated graphs, tables, and two synthesis metrics, one for measurements by criminal category and the other for measurements by municipalities. The results pointed out that the five most criminally impacted municipalities in 2017 and 2018 had very high indexes, in comparison to 2019, where most municipalities got lower indexes. The most violent municipalities are characterized by poor socioeconomic factors, like basic sanitation, Regional Human Development Index, and occupation. Geographical visualization of the collected results was also generated.

Using a criminal data set for the years between 2014 and 2019 in Denver, United States, (Ratul, 2020) conducted a comparative study of machine learning algorithms. Before the application of the algorithms, cleaning, reduction, integration, conversion, shuffling, normalization, sampling, and feature selection steps were executed. Ten modeling techniques were evaluated with three different strategies: train-test split, cross-validation, and paired t-test. The best model was an ensemble algorithm, with an overall accuracy of 90% for 15 different types of crimes, the set with the highest amount of classes.

An empirical comparison between a proposed ensemble model and individual classification algorithms was accomplished by (Kshatri, et al., 2021) using criminal offense data sets. The collected data consists of murder, rape, robbery, and other violent crimes in India, between 2001 and 2015, which were submitted to extraction, transformation, and reduction steps. Among the individual modeling techniques, the bagging ensemble technique was the most efficient, achieving 95.55% of accuracy in the test set. The proposed model, called staking ensemble, achieved 99.5%, proving the efficiency of ensemble algorithms over individual models.

Aiming to understand the relationship between criminal events against women, (Vinholes, 2019) proposed a solution using database knowledge discovery. After being processed through cleaning, integration, resizing, and transformation steps, two data sets of criminal occurrences against the feminine population in the Brazilian state of Rio Grande do Sul were combined and submitted to the Apriori algorithm. The generated associative rules indicate the following results: 61% of all periods with low incidences of lesions, there are low occurrences of threat; 52% of all periods with high incidences of rape, there are low occurrences of threat; and 50% of all periods with high incidences of bodily injuries, there is low occurrences of threats.

Another comparative analysis between machine learning algorithms use data from different sources to predict trends and the number of occurrences of crimes by type and geographic region, as seen in (Castro, 2020). Theft and robbery data for the Brazilian state of Minas Gerais were collected and submitted to the selection, transformation, and feature analysis steps, and then used in different machine learning algorithms. The Long-Short Term Memory (LSTM) model achieved a higher accuracy, approximately 91%.

In order to accomplish an evaluation of governmental public security data, two experiments were conducted using databases in both spheres, national and within the state of Minas Gerais, accordingly to (Prado, 2020). A centralized architecture of ETL (Extraction, Transform, Load) and data mining generated two applications for each scope mentioned in the research problem, along with associative rules extraction between geographic regions with types of crime and targets of robbery.

In (Costa, 2020), an unsupervised approach of clustering between the 185 municipalities of the Brazilian state of Pernambuco was developed using criminal variables. First, the dimensionality reduction technique of Principal Component Analysis was applied, followed by the use of the k-means algorithm, which allowed to group the municipalities for different values of k, where $k = 26$ achieved satisfactory results, which is explainable with the coincidental number of 26 Integrating Areas of Public Security in the state. Then, the three most common types of crime for each cluster were listed, using the moving average indexes for each crime within each municipality.

Lastly, situated in Salvador, Brazil, the work described in (Amorim & Pereira, 2019) describes the categorization of criminal events using different machine learning algorithms. The data set consisted of 465,376 samples and 17 attributes, distributed in 381 classes of crime. The data preparation step was able to remove punctuation marks, remove stop words, treat missing values, remove diacritic marks, and lowercase all text. The proposed algorithms were evaluated with the Hold-out approach, where both C4.5 and Rippier algorithms achieved an accuracy of 99%.

3. Methodology

CRISP-DM (CRoss-Industry Standard Process for Data Mining) is an industry-driven approach to guide data mining process models, detailing its life cycle and the associated tasks of each phase. It is commonly divided into six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Although most guides describe the operation of each phase, CRISP-DM is designed to be flexible and easily customized, where some phases might be reallocated, considered less important than others, or simply omitted, accordingly to the general purpose of the project. Figure 1 illustrates how CRISP-DM was redesigned to fit the objectives of the current paper, focusing mainly on the data preparation, modeling, and evaluation steps.

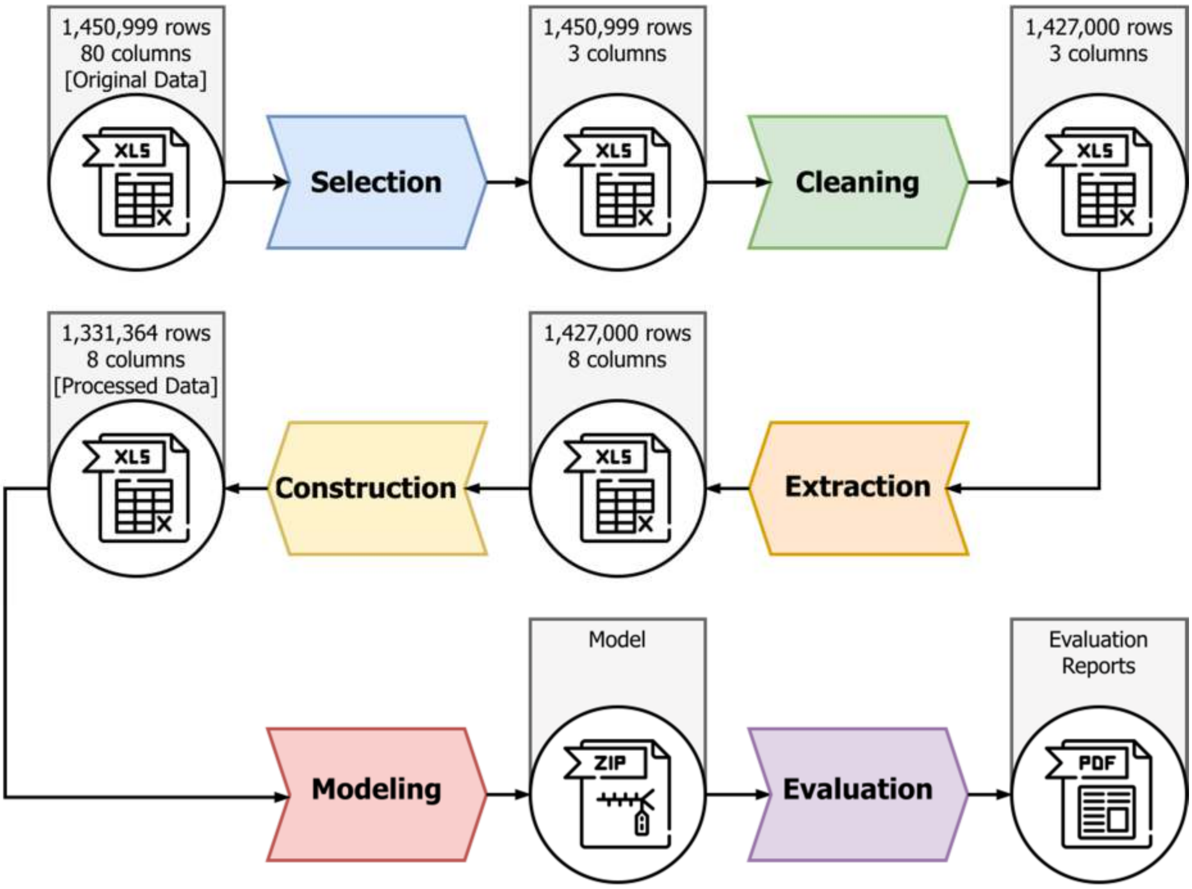

Figure 1. Phases of the CRISP-DM reference model adapted to the development of the classifier.

Under those circumstances, the development of the classifier consists of the six tasks derived from the CRISP-DM model process, executed in the aforementioned sequence. The selection of these tasks aided in the construction of the respective software, mainly developed using the programming language Python, as it is one of the leading technologies related to data manipulation and data mining tasks.

Concerning the gathering of data used in the analysis, the Assistant Secretariat for Intelligence and Criminal Analysis (SIAC) made available a collection of the police records for the years 2019, 2020, and 2021. Stored in tabular files, the combined data contains 1,450,999 samples and 80 attributes, mostly related to record identification, time, location, and *modus operandi* settings, general data of victims and prosecutors, the descriptive report of the event, and criminal typology, such as registries given by the respective police station and the consolidation of the event, the most accurate labeling of the record within SIAC's database, which is used for the statistical purposes of the secretariat. As a result of the secretariat demands for an automatic confirmation tool of the main criminal events labeling, to aid the reading of the massive volume of daily records in a reduced amount of time and effort, along with the research's need for an accurate and descriptive set of labels, the latter attribute was chosen as the target column, which guided the decisions of the following step.

The selection step decides on the data to be used for analysis, based on the relevance to the project objectives, along with the selection of attributes (columns) as well as the selection of records (rows). Only three columns were judged as important for the development of the classifier: “nro_bop”, the unique identifier of each record; “report”, the complete description of the event; and “consolidated”, the label given by criminal analysts after the reading of a report.

The cleaner step raises the data quality to make the application of data mining techniques viable. It consists of the selection of clean subsets of data, the removal of duplicated records, the removal of outliers - defined in terms of the report's length -, and the removal of classes with too few amount of samples to successfully extract knowledge.

The extraction step applies feature engineering techniques to discover meaningful attributes in existing data. The main indication of qualitative meaningfulness of the reports is the quantization of its report elements, such as words, characters, or sentences, along with their descriptive statistical metrics.

The construction step concludes data preparation, producing derived attributes, entire new records, or transformed values for existing attributes, and generating the final form of the data to be submitted to the modeling step. The textual column of reports was lowered, striped of HTML tags (originated on the source database), multiple spaces, punctuation, and diacritic marks, together with sensitive information encoding, where personal data about victims and prosecutors were substituted by a constant label, and lemmatization, transforming each word to its root form. The consolidated labels were also modified, removing unwanted classes, such as “ALREADY LAUNCHED” (duplicate of an existent record with a different identifier) and “NOT INFORMED” (for defective reports), along with a grouping of similar classes, for instance, “ROBBERY” and “VEHICLE ROBBERY” became a single class, under the name of the former. Figure 2 illustrates a ranking of the 25 most frequent labels of the processed data.

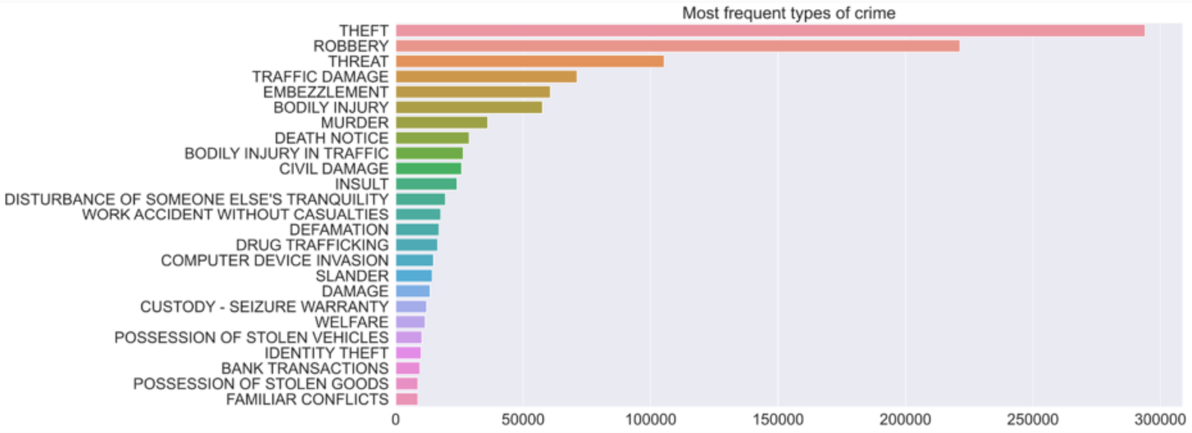

Figure 2. The 25 most frequent target labels of the data set.

The modeling step is responsible for executing the modeling tool on the processing data set to create predictive models. A tokenizer object was constructed, to capture all words present in the data set and map each of them to a unique index. Then the text was transformed into sequences of numbers and limited to a maximum sequence length. As for the consolidated target labels, the One-Hot Encoding technique transformed each of them into a unique number. The 1,331,364 instances that came from the construction step were divided into 3 subsets: train (70%), validation (15%), and test (15%). Finally, the proposed model, illustrated in Figure 3, was fit to the training data, starting the capture of knowledge in the database.

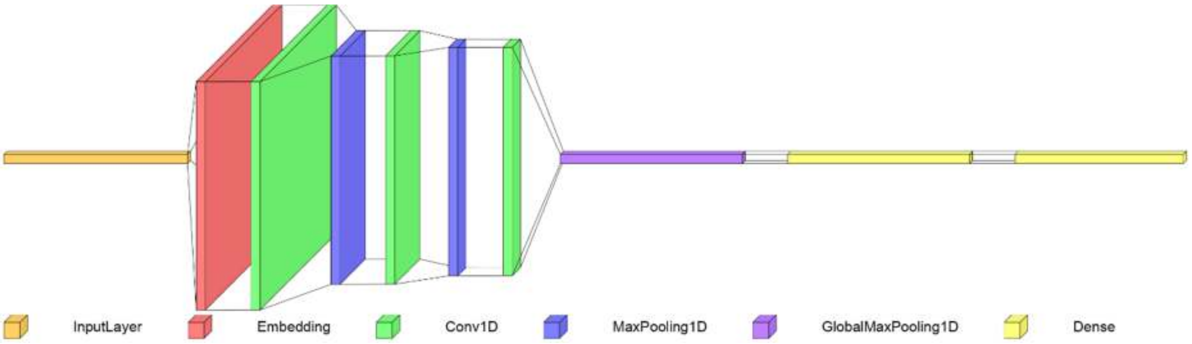

Figure 3. Layers of the proposed Deep Learning model.

As for the choice of the layers used, (Kim, 2014) proposes the use of Convolutional Neural Networks (CNN) in the classification of sequences of text, where a convolution operation applied by a filter over local chunks of data in a word vector generates a feature map of patterns, whose maximum value can be passed to a fully connected activation layer, producing the output vectors for the classification purposes. In the same manner, (Chollet, 2018) discusses how CNN architectures are computationally cheaper and quite competitive in regard to classification performance, in comparison to models usually found in text classification schemes, such as Recurrent Neural Networks (RNN). The latter author also highlights the advantages of using word embeddings over traditional encoding techniques of categorical data (such as One-Hot Encoding), creating dense word vectors that distribute the words over a vector space using a semantic coherent approach, where similar words are close to each other. Thus, the proposed architecture is composed of the following schema:

- An embedding layer reshaped the input data into a 3D dense tensor, capable of understanding the semantic relationship between the words, 500-dimensional and limited to 500 words per sentence. Given the nature of the proposed problem, where a great number of words are exclusive of the public security environment, the embedding layer was trained from scratch, rather than using pre-trained embeddings, such as GloVe and word2vec.
- A sequence of convolution layers over a single spatial dimension and max pooling layers were able to extract feature maps in the sliding window of attributes of the embedding word vector, along with dimensionality reduction operations.
- A couple of fully-connected dense layers activate the neurons corresponding to each of the 463 labels, generating the probability distribution of predictions.

The learning algorithm was set to run during a single execution of 30 epochs, with an option to interrupt the execution after 10 epochs without improvements on the validation loss. As shown in Figure 4, the best value of validation loss was achieved in the second epoch ($val\_loss = 0.7042$, $val\_accuracy = 0.8105$), and the learning algorithm was interrupted in the 12th epoch.

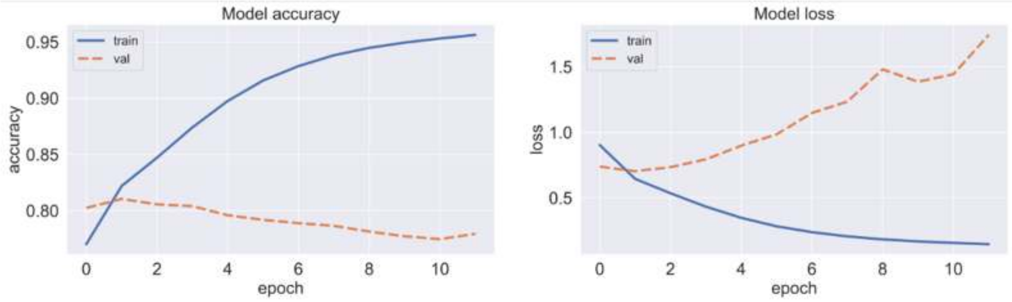

Figure 4. Accuracy and loss evolution through the training epochs.

With the generation of the best model, the evaluation step is responsible for the assessment of the degree to which the model meets the project objectives. The results of the evaluation will be analyzed in the following section.

4. Evaluation

4.1. Train test split evaluation

The proposed model was evaluated using the test set of 199,705 samples, achieving an overall accuracy of 77.89%. A 463 × 463 confusion matrix describes the behavior of the model in the test set, indicating the correct predictions and mispredictions. Given the dimension of the confusion matrix and the unbalanced nature of the test data, the exposition of its results becomes quite polluted. Figure 5 illustrates a reduced view of the confusion matrix, calculating the individual metrics (True Positives, True Negatives, False Positives, and False Negatives) for 9 types of crimes of interest.

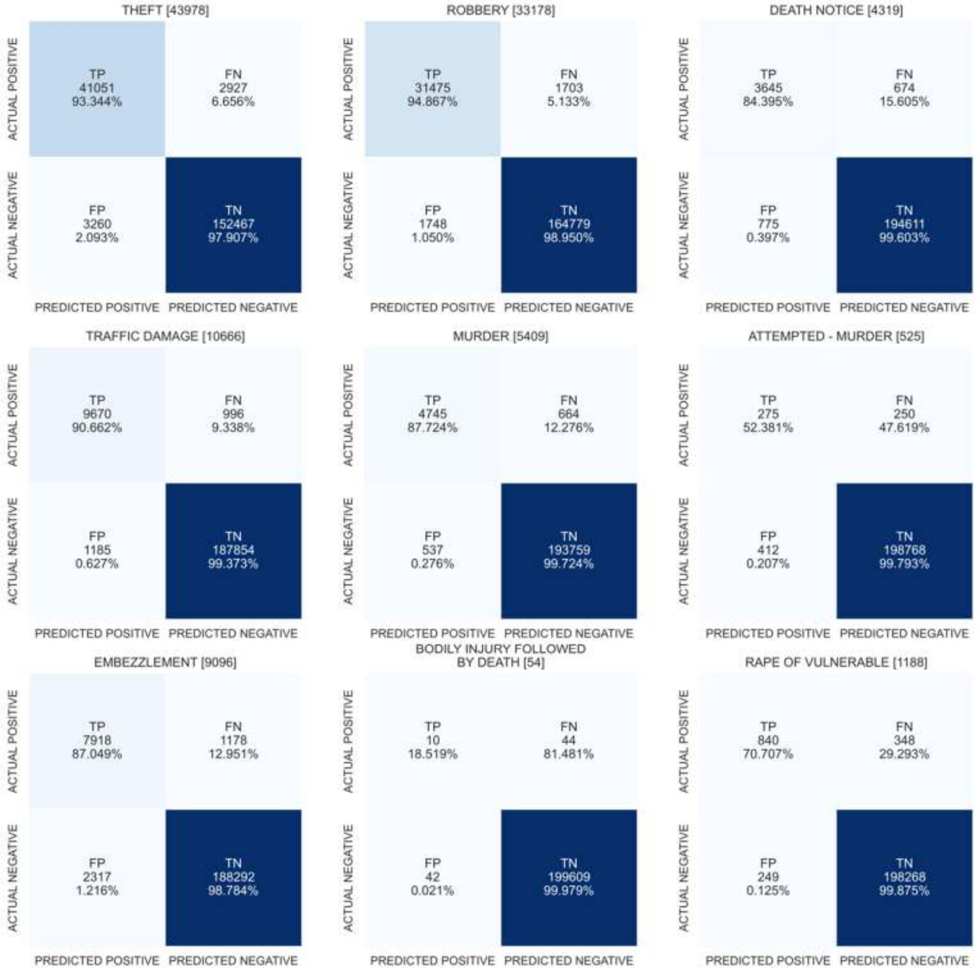

Figure 5. Confusion Matrices for the labels of interest.

A group of evaluation metrics were collected from the general confusion matrix, and are exposed in Table 1, for a subset of the labels judged as the most interesting for the paper. Robbery, theft, murder, and traffic damage are the labels with higher predictive performance, as indicated by their values of precision (hits among true predictions) and sensitivity (hits among relevant predictions) being higher than 85%. The Matthews Correlation Coefficient (MCC) is a useful metric to measure multi-classification problems in unbalanced data sets, providing a uniform calculation between all four indexes of the individual confusion matrices, indicating that the same four aforementioned labels present satisfactory performance.

Table 1. Evaluation metrics for the labels of interest.

Label	Accuracy	Precision	Sensitivity	F1-score	MCC
THEFT	0.969019	0.926429	0.933444	0.929923	0.910048
ROBBERY	0.98272	0.947386	0.948671	0.948028	0.937665
DEATH NOTICE	0.992744	0.824661	0.843945	0.834192	0.830541
TRAFFIC DAMAGE	0.989079	0.890834	0.906619	0.898657	0.892925
MURDER	0.993986	0.898334	0.877242	0.887663	0.884639
ATTEMPTED - MURDER	0.996685	0.400291	0.52381	0.453795	0.456282
EMBEZZLEMENT	0.982499	0.77362	0.870493	0.819202	0.811597
BODILY INJURY FOLLOWED BY DEATH	0.999569	0.192308	0.185185	0.188679	0.188498
RAPE OF VULNERABLE	0.997011	0.77135	0.707071	0.737813	0.737017

An alternative visualization of the confusion matrix is illustrated by the chord diagram in Figure 6. A chord diagram is commonly used to visualize many-to-many relationships, where each entity is expressed by the length of an arch of circumference, and the relations are ribbons linked across the circle. The presented diagram expresses mispredictions across labels, such as a higher amount of robbery records predicted as theft - which is expected given the set of common words between these reports -, and vice-versa, for example. It is important to notice that this diagram does not show a general overview of the predictions (only labels with more than 500 true samples were selected), but is the closest representation achieved for the current number of classes.

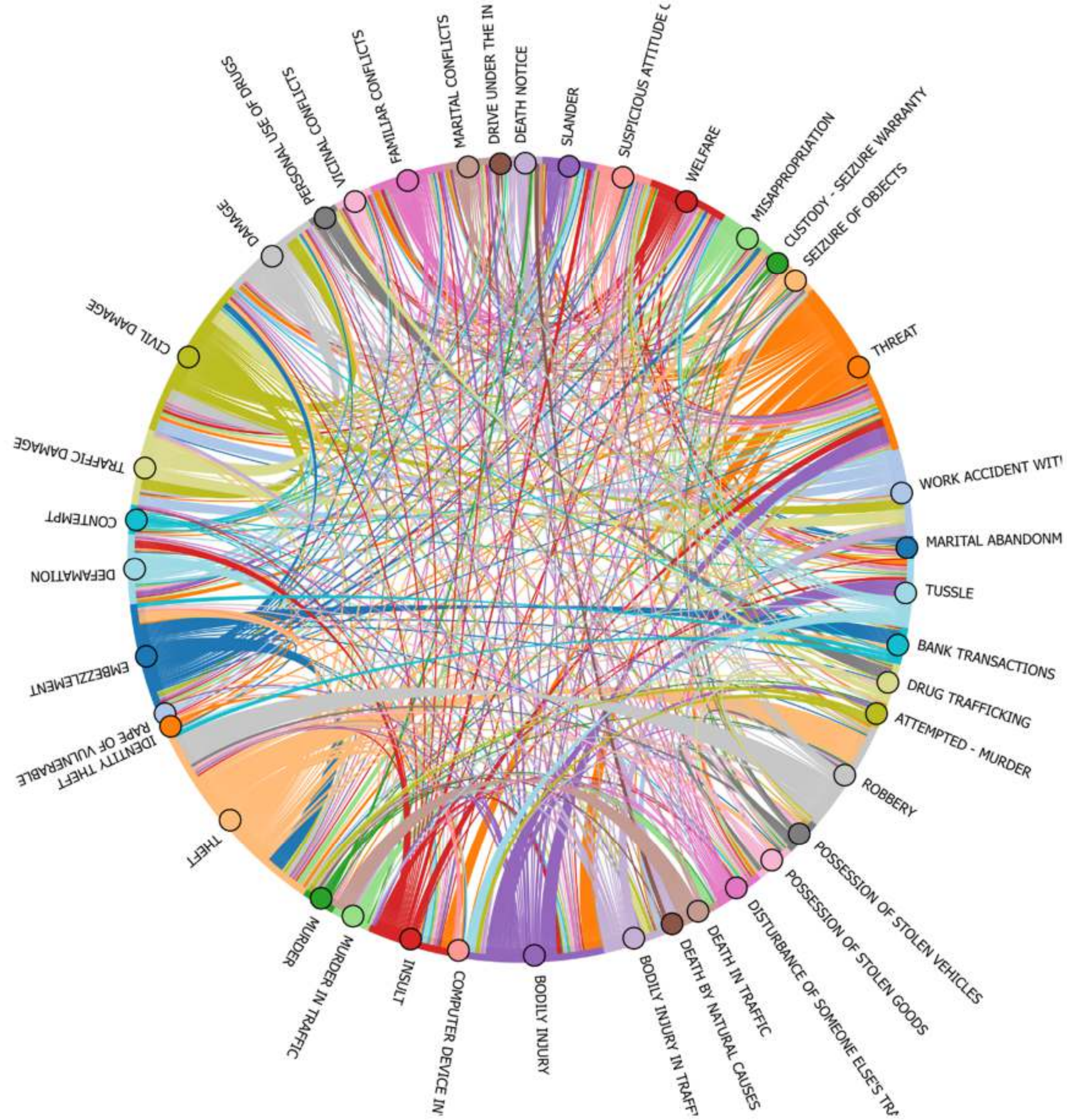

Figure 6. General overview of predictions between labels of interest.

An alternative visualization for individual types of crime is shown in Figure 7. A Sankey diagram represents flows of data from a source node to a target node. Thus, each diagram illustrates the flows of predictions of a certain class. This visualization highlights not only the proportion of reports that were correctly classified, but also the classes that usually come as mispredictions, and their likeness to have common words with the expected class. For instance, murder is likely associated with violent crimes against a person, death notice is likely associated with deaths by natural causes and accidents, theft with crimes against the patrimonial, likewise to rape of vulnerable being mispredicted with crimes against sexual dignity.

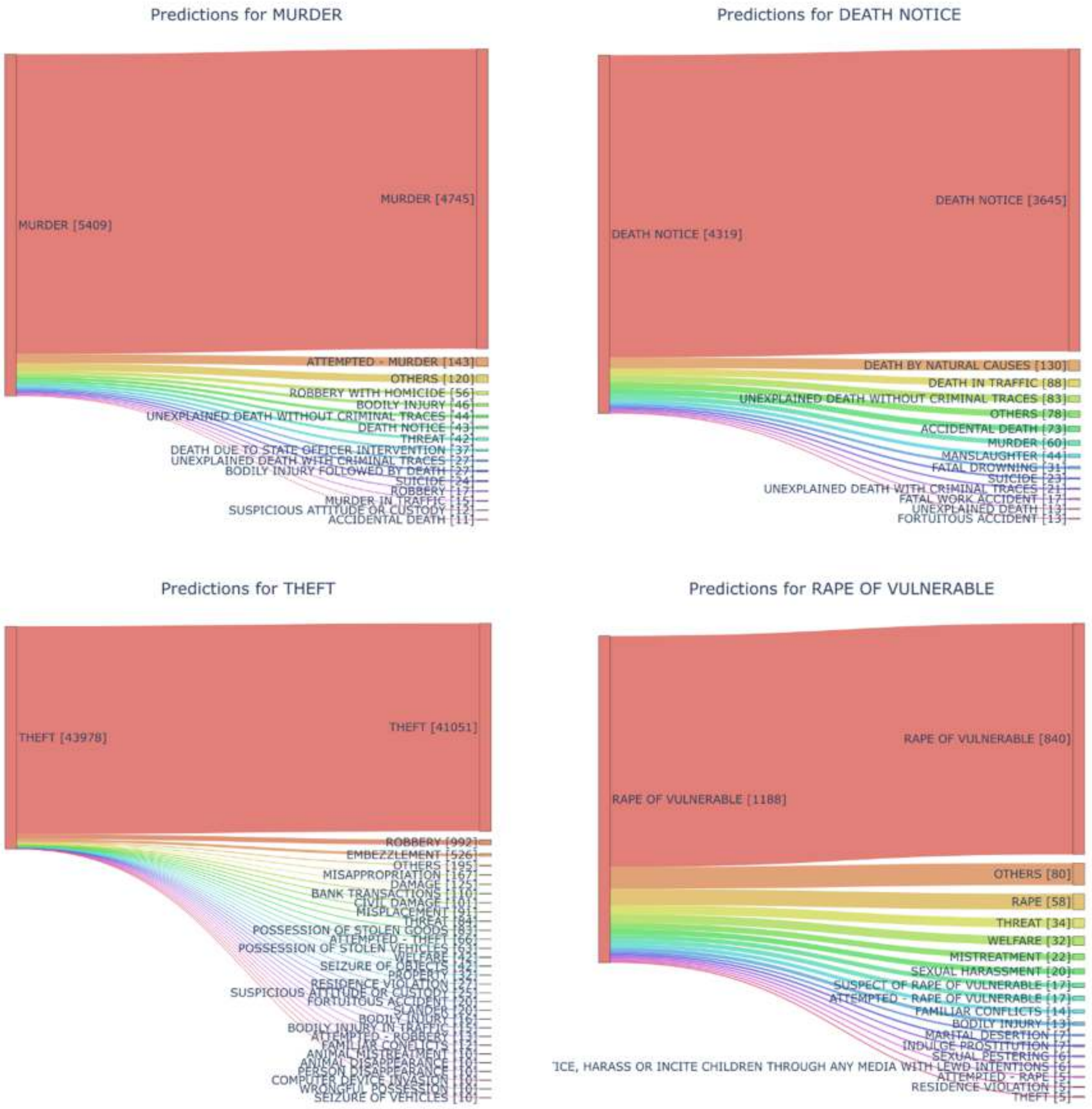

Figure 7. Flow of predictions for four labels of interest.

4.2. Production evaluation

Following the fulfillment of the research's methodology and evaluation, an agreement with SIAC's statistical and criminal analysis department has been established, using the proposed model as part of a tool to automatically label the daily reports.

The tool was incorporated into the department's routine in the generation of the consolidated column, the confirmation of the labeled data originated in the public security systems of the State of Pará against the actual textual description of the report. The original label is not always consistent as it is attributed at the report registration without a complex analysis of the event, which might compromise the statistical demands of the secretariat. The department allocates a group of 15 criminal analysts to read and tag the reports to a specific label based on their empirical knowledge of the Brazilian legislation, which is usually adapted to fit statistical purposes, and the collective judgment of the analysts, for complex cases. Currently, this analysis is produced over a selected set of reports, mainly focusing on violent crimes (murder, robbery, rape, etc.), due to the huge volume of daily data generated on the police stations. The reports that are not in the selected reading set receive a consolidated label adapted from the original label. This methodology indicates that approximately 10% of the original labels are not compatible with the consolidated label, an error that is the main target of the allocation of efforts to improve the quality of the data.

Under those circumstances, the proposed model was integrated into a processing tool with the objective of labeling the reports prior to the analysis of the departments' experts. Only the reading set is being used in the current approach, given the low precision of the model for most of the samples not read by analysts. A comparison of the predictions of the classifier and the original label of the police stations indicates the mispredictions, the group of reports that need to be read by analysts. As for the correct predictions, most of the reports are set with the label generated by the classifier, except for violent deaths that require a deep human analysis (murder, manslaughter, death notice, etc.). Therefore, the model assesses a quantitative amount of consolidated samples and highlights the records that need attention.

A production test set was selected for the period between March 16 and October 5, 2022, for 266,595 samples. An analysis of the direct comparison (hits and misses based on string matching) between the predicted and expected labels was conducted, considering the expected labels of the police stations (PS) and the consolidated labels generated in the secretariat (CS). Figure shows the percentage of hits in the comparison between predictions and expected labels, for different categories of reports: the "raw" set ($PS = 63.1\%$ and $CS = 73.5\%$) considers all available samples, without restriction;

- the "learned" ($PS = 66.4\%$ and $CS = 77.4\%$) set considers only the samples whose consolidated label was learned by the classifier, among the 463 selected for training;
- the "read" set ($PS = 61.5\%$ and $CS = 75.1\%$) considers only the samples read by the secretariat, the more accurate labels given by the analysts. It is used in the daily predictions of the classifier, as its aftermost human analysis allows its evaluation;
- the "unread" set ($PS = 65.1\%$ and $CS = 71.4\%$) considers only the samples not read by the secretariat, the least accurate labels adapted from the original label of the police stations;
- the "read_learned" set ($PS = 66.7\%$ and $CS = 81.4\%$) considers only the samples read by the secretariat whose consolidated label was learned by the classifier;
- the "unread_learned" set ($PS = 63.1\%$ and $CS = 73.5\%$) considers only the samples not read by the secretariat whose consolidated label was learned by the classifier;

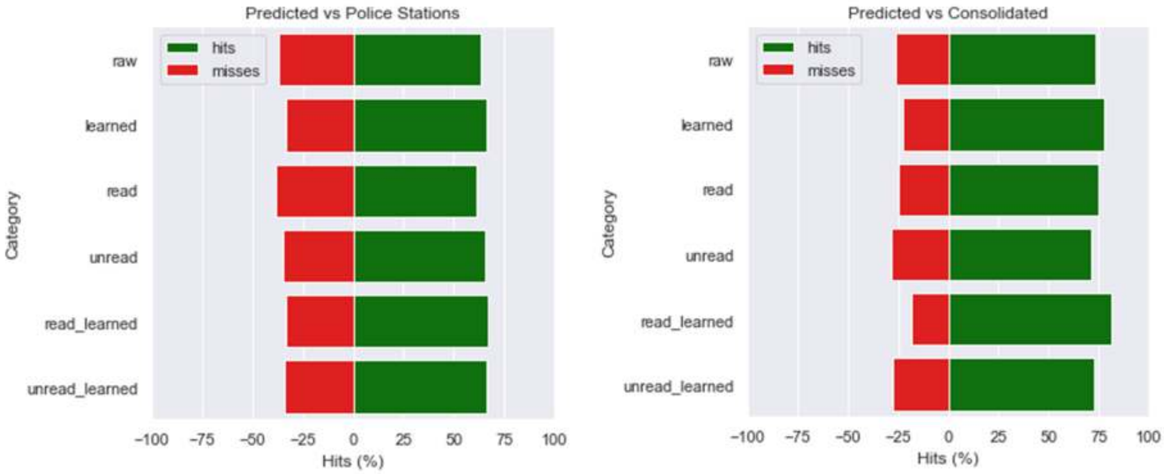

Figure 8. Proportion of hits of the comparison between predictions and expected labels.

The results indicate that the classifier can reduce the reading efforts of labeling the reports by at least 61.5%, the proportion of samples read by analysts with labels not necessarily learned by the classifier in comparison with the classes available before human analysis. This percentage includes violent death classes, which are always double-checked by analysts. As for the read samples with consolidated labels learned by the classifier, the high amount of hits indicates the importance of curating the set of labels to be learned, one of the main challenges of the proposed problem, given the unbalanced dataset.

Both sets of samples not read by analysts take into account most of the rare classes that are not normalized, are not prioritized in the criminal analysis process, were registered before 2019 (the first of the triennial cut of the dataset), or are not used anymore by the secretariat. An analysis provided by the same analysts over the missed samples can indicate a precise amount of samples whose labels could be adjusted by the classifier, generating more accurate classes for the overlooked unread labels to be used in further classification models.

The implementation of the classifier brought not only qualitative improvements in the data processed in the secretariat, by reallocating the analyst's workload to the reports more difficult to label but also a reduction in the time of generation of the daily reports that the public administration demand. The statistical department mentioned that, prior to the use of the tool, the reports were released up to a day after the collection of the records, in comparison to the current average release, an average of 6 hours after the start of the daily production. The actual impact of the implementation of the tool is yet to be measured and evaluated.

5. Conclusion

The current paper presented the construction of a data mining tool used to extract knowledge from a police report database through the application of a supervised classification model. The proposed architecture was based on CRISP-DM, a standardized process model usually applied to data mining tasks. The classification model achieved an overall accuracy of 78%, a result mostly impacted by the high amount of target classes in the classification problem and the fact that the counting of learning examples was unbalanced, hindering the performance for rare classes. The model has the potential to be used by the statistical agencies related to public safety in the state of Pará, as it can automate the process of confirmation of certain events within the description of the records, along with an unbiased and deterministic verdict. The impact of the application of the classifier in such agencies is yet to be measured and can be the motivation for further research. Other public safety agencies can also benefit from the use of the proposed model since the collected knowledge relied only on the collection of patterns extracted from a textual description of labeled events, thus the generated tool can be used from different perspectives, with the necessary modifications, as it is completely independent of the legislative environment where it will be applied.

Possible improvements to the classifier include the adoption of more steps of the CRISP-DM model, a fine decision of which target labels are to be learned, the balancing of the learning examples for all classes, the testing of different machine learning algorithms, and the application of hyper-parameter optimization algorithms. Regarding the pre-processing steps, research on Named-Entity Recognition (NER) for the highlighting of socioeconomic attributes inside the reports, feature engineering, and statistical significance of samples can be explored to improve the quality of the data to be processed. An effort of integration between the research fields of law and artificial intelligence can be reached to perform the application of knowledge specific to the Brazilian law codes in the discovery of the data patterns deterministic to the prediction of the consolidated events.

Bibliography

Aldossari, B. S., Alqahtani, F. M., Alshahrani, N. S., Alhammam, M. M., Alzamanan, R. M., Aslam, N., & Irfanullah. (2020). A Comparative Study of Decision Tree and Naive Bayes Machine Learning Model for Crime Category Prediction in Chicago. *Proceedings of 2020 the 6th International Conference on Computing and Data Engineering* (pp. 34–38). Sanya, China: Association for Computing Machinery.

Amorim, M. d., & Pereira, J. R. (2019). Tipificação de ocorrências policiais utilizando machine learning. (p. 50). Salvador: Universidade Católica do Salvador.

Assad, F. J., & Chagas, J. F. (2019). Análise Preditiva de Manchas Criminais no Estado de São Paulo. (p. 93). Niterói: Universidade Federal Fluminense.

Brasil. (2022, March 16). *O Sinesp*. Retrieved from Ministério da Justiça e Segurança Pública: https://www.gov.br/mj/pt-br/assuntos/sua-seguranca/seguranca-publica/sinesp-1/

Castro, U. R. (2020). Explorando aprendizagem supervisionada em dados heterogêneos para predição de crimes. (p. 85). Belo Horizonte: Pontifícia Universidade Católica de Minas Gerais.

Chen, P., & Kurland, J. (2018). Time, Place, and Modus Operandi: A Simple Apriori Algorithm Experiment for Crime Pattern Detection. *2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA)*, (pp. 1-3).

Chiew, L. S., Amerudin, S., & Yusof, Z. M. (2020). A spatial analysis of the relationship between socio-demographic characteristics with burglar behaviours on burglary crime. *IOP Conference Series: Earth and Environmental Science*, 012050.

Chollet, F. (2018). *Deep Learning with Python.* Manning Shelter Island.

Costa, J. C. (2020). Identificação de municípios pernambucanos para recomendação de políticas de segurança pública utilizando uma técnica de clusterização. (p. 70). Caruaru: Universidade Federal de Pernambuco.

Furtado, L., & Souza, A. (2019). Uso de Dados Provenientes de Rede Social e Técnica de Mineração de Dados para Classificar Crimes em Belém-PA. *The Academic Society Journal*, 121-134.

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1746-1751). Doha, Qatar: Association for Computational Linguistics.

Kshatri, S. S., Singh, D., Narain, B., Bhatia, S., Quasim, M. T., & Sinha, G. R. (2021). An Empirical Analysis of Machine Learning Algorithms for Crime Prediction Using Stacked Generalization: An Ensemble Approach. *IEEE Access*, 67488-67500.

Prado, K. H. (2020). Data science aplicada à análise criminal baseada nos dados abertos governamentais do Brasil. (p. 146). São Cristóvão: Universidade Federal de Sergipe.

Qazi, N., & Wong, B. W. (2019). An interactive human centered data science approach towards crime pattern analysis. *Information Processing & Management*, 102066.

Ratul, M. A. (2020). A Comparative Study on Crime in Denver City Based on Machine Learning and Data Mining. *CoRR*.

Regateiro, H. A., Ramos, E. M., Souza, J. G., & Mello, C. M. (2021). Assessment of crime in the State of Pará. *Research, Society and Development*, e10010313088.

Souza, D. C. (1972). *Introdução à Ciência do Direito.* FGV.

Souza, J. R. (2018). Utilização de aprendizagem de máquina na predição de crimes. (p. 54). Niterói: Universidade Federal Fluminense.

Trindade, E. A. (2019). Homicídios na Região Metropolitana de Belém: práticas para contenção e vulnerabilidades. (p. 155). Belém: Universidade Federal do Pará.

Vargas, W. A. (2019). Data science & segurança pública: padrões estatísticos sobre as ocorrências de flagrantes em roubo de celular na cidade de São Paulo. (p. 52). São Paulo: Fundação Getúlio Vargas.

Vinholes, T. V. (2019). Descoberta de Conhecimento em Banco de Dados Relacionados à Violência Contra a Mulher. *Anais do Computer on the Beach*, Anais do Computer on the Beach.