A Supervised Classifier for Police Reports at the State of Pará, Brazil

Helder Matos[1], **Samara Souza**[1], **Reginaldo Santos**[1],
João Weyl Costa[2], **Cleyton Costa**[3]

[1] Faculdade de Computação
Universidade Federal do Pará (UFPA)
Av. Augusto Correa 01, 66075-090 – Belém – PA – Brasil

[2]Faculdade de Engenharia da Computação e Telecomunicações
Universidade Federal do Pará (UFPA)
Av. Augusto Correa 01, 66075-090 – Belém – PA – Brasil

[3]Programa de Pós-Graduação em Segurança Pública
Universidade Federal do Pará (UFPA)
Av. Augusto Correa 01, 66075-090 – Belém – PA – Brasil

{helder.matos,samara.souza}@icen.ufpa.br

{regicsf, jweyl}@ufpa.br, cleyton.costa@ifch.ufpa.br

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.*

1. Introduction

The National System of Public Safety Information (SINESP) is the main infrastructure of data and criminal information in Brazil, providing a secure and standardized communication system between members of the Unified Public Security System (SUSP) [Brasil 2019]. Each state of the federation is responsible for the processing of data extracted from police stations throughout the country. This integration requires a considerable amount of effort, in order to consolidate a massive volume of police records.

The application of data science and data mining tools became a trend to be observed in the digitization of processes in public sectors, including public safety [de Vargas 2019, da Silva Amorim and Pereira 2019, de Castro 2020, Ratul 2020, Kshatri et al. 2021]. Moreover, public safety can acquire benefits generated by automatic tools of knowledge extraction in databases, including classification of police reports, the detection of associative rules, information visualization, business intelligence, and extraction and generation of hidden attributes in public databases.

The current paper describes the development of a supervised classifier of police reports, using data from the Assistant Secretariat for Intelligence and Criminal Analysis (SIAC), in the state of Pará, Brazil. It processes police reports and generates the prediction of a type of event. The resulting model is used to accelerate the statistical processes of the secretariat, along with an automatized qualitative analysis of a huge amount of data.

2. 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. Among the many phases described in the CRISP-DM documentation, six of them were chosen as relevant for the research. The tool was mainly developed using the Python programming language, one of the leading technologies related to data mining tasks, with a wide range of tools related to Natural Language Processing.

SIAC made available data of the police records for the years 2019, 2020, and 2021, composed of 1,450,999 samples and 80 attributes, including: the record identification; the descriptive report of the event; and the consolidation of the event, an accurate labeling used for the statistical purposes of the secretariat. As a result of the secretariat's need for an automatic reading of the criminal events to aid the processing of the 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.

The **selection** step decides on the data relevant for the analysis, through the choice of attributes and records important for the development of the project: "nro_bop", the unique identifier of each record; "report", the description of the event; and "consolidated", the label given by a criminal analysts. The **cleaning** step raises data quality through the selection of clean subsets of data, the removal of duplicated records, the removal of outliers (reports with short length of text), and the removal of classes with too few samples to successfully extract knowledge. The **extraction** step applies feature engineering techniques to discover meaningful attributes in existing data. The main indicative of qualitative meaningfulness of the reports is the quantization of its textual elements, such as words, characters, or sentences. The **construction** step produces derived attributes, entire new records, or transformed values for existing attributes. The textual column of reports was lowered, striped of HTML tags originated on the source database, multiple spaces, punctuation, and diacritic marks, along with sensitive information encoding and lemmatization. The labels were modified, removing unwanted instances and grouping similar criminal classes, which generated the 463 labels of the problem. 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 the words present in the data set and map each of them to a unique index. Then the text was transformed into sequences of numbers 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 of the construction step were divided into 3 subsets: train (70%), validation (15%), and test (15%). Finally, the proposed model was fit to the training data.

As for the choice of the layers, [Kim 2014] proposes the use of Convolutional Neural Networks (CNN) in the classification of texts, 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. In the same manner, [Chollet 2018] discusses how CNNs are computationally cheaper and competitive in comparison to other text classification schemes, and explains the advantages of using word embeddings over traditional encoding of categorical data, creating dense vectors that distribute the words over a vector space using a semantic approach. Thus, the proposed model is based on such CNNs.

The learning algorithm ran during a single execution of 30 epochs, set to be interrupted after 10 epochs without improvements on the validation loss. Figure 1 shows that the best value of validation loss was achieved in the 2nd epoch ($val\_loss = 0.7042$, $val\_acc = 0.8105$), and the learning algorithm was interrupted in the 12th epoch.

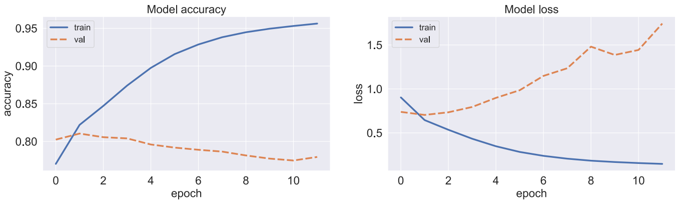


Figure 1. Accuracy and loss evolution through the training epochs.

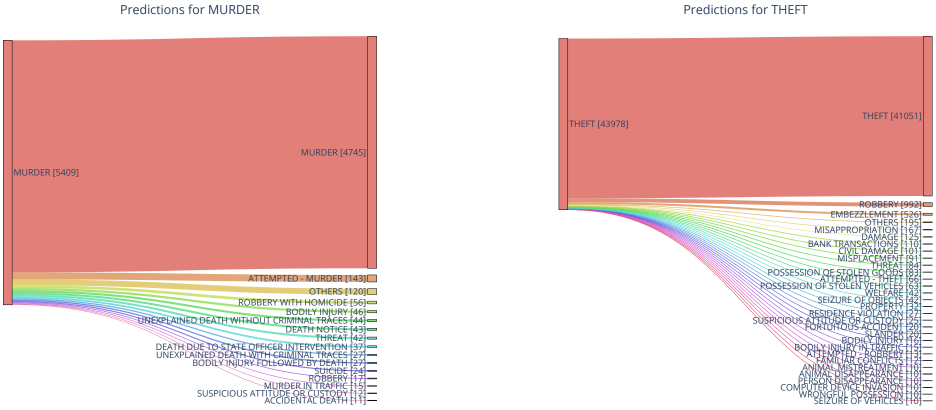
3. Results and discussion

The **evaluation** step assesses the degree to which the model meets the objectives. The best model was tested on a set of 199,705 samples, with an overall accuracy of 77.89%. The evaluation metrics for six labels are exposed in Table 1, where four of them achieved over 85% of precision and sensitivity (robbery, theft, traffic damage, and murder). The Matthews Correlation Coefficient (MCC) is a proper metric for multi-classification problems in unbalanced data, providing a uniform calculation between all indexes of the confusion matrix, where the same four labels had better results.

Table 1. Evaluation metrics for the labels of interest.

Label	Accuracy	Precision	Sensitivity	F1-score	MCC
THEFT	0.9690	0.9264	0.9334	0.9299	0.9100
ROBBERY	0.9827	0.9473	0.9486	0.9480	0.9376
DEATH NOTICE	0.9927	0.8246	0.8439	0.8341	0.8305
TRAFFIC DAMAGE	0.9890	0.8908	0.9066	0.8986	0.8929
MURDER	0.9939	0.8983	0.8772	0.8876	0.8846
RAPE OF VULNERABLE	0.9970	0.7713	0.7070	0.7378	0.7370

Figure 2 shows the flows of predictions of a certain class, highlighting not only the proportion of hits, but also the mispredictions and their likeness to have common words with the expected class. For instance, murder is likely associated with violent crimes against a person, and theft with crimes against the patrimonial.



(a) Murder.

(b) Theft.

Figure 2. Flow of predictions for two labels of interest.

4. Conclusion

The paper described the construction of a data mining tool used to extract knowledge from police reports using a supervised classification model. It achieved an overall accuracy of 78%, a result hindered by the high amount of classes and the use of an unbalanced data set. It has the potential to be used by 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, as the process relied only on the collection of patterns extracted from a textual description of labeled events, completely independent of the legislative environment where it will be applied.

The classifier can be improved with a 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. 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.

References

Brasil (2019). O Sinesp. Date of access: March 16, 2022.

Chollet, F. (2018). *Deep Learning with Python*. Manning Shelter Island, 1st edition.

da Silva Amorim, M. and Pereira, J. R. S. (2019). Tipificação de ocorrências policiais utilizando machine learning. p. 50.

de Castro, U. R. M. (2020). Explorando aprendizagem supervisionada em dados heterogêneos para predição de crimes. p. 85.

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

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

Kshatri, S. S., Singh, D., Narain, B., Bhatia, S., Quasim, M. T., and Sinha, G. R. (2021). An Empirical Analysis of Machine Learning Algorithms for Crime Prediction Using Stacked Generalization: An Ensemble Approach. *IEEE Access*, 9:67488–67500.

Ratul, M. A. R. (2020). A Comparative Study on Crime in Denver City Based on Machine Learning and Data Mining. *CoRR*, abs/2001.02802.