ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 8, August 2024

Developing a Framework for Social Data Analysis: A Study Based on Criminal Records from São Paulo

[1] Diego De Castro Rodrigues, [2] Marcelo B. Nery, [3] Sergio Adorno

[1] [2] [3] University of São Paulo – USP

Corresponding Author Email: [1] diego.rodrigues@ifto.edu.br, [2] marcelo@usp.edu.br, [3] sergio@usp.edu.br

***Abstract*— This project aims to establish a framework for social data analysis, with a particular emphasis on leveraging criminal records and applying descriptive computational techniques such as associative algorithms and decision tree rule extraction. The methods and tools discussed in this work will enable the identification of patterns, providing a guided means to recognize similarities between recurring situations in the social sphere, utilizing descriptive techniques and data visualization. The study area has been defined as the city of São Paulo, focusing on the structuring of social data and the quality of the information. A set of tools will be validated, including the use of databases and visualization tools for the results. Among the main deliverables of the project are the discoveries made during the research phase. The effectiveness and utility of the results will rely on studies involving real data, validated both by domain experts and through the identification and comparison of the patterns found in this study with.**

***Index Terms*— Social Data Analysis, criminal records, computational techniques, Data Mining, Big Data.**

I. INTRODUCTION

Studies on crime are of great importance to understand the factors that lead to the occurrence of crimes and to develop effective public policies to prevent them. In the city of São Paulo, one of Brazil's main urban centers, crime is a complex and multifaceted problem that affects the lives of citizens and the city's economy.

Several authors have been dedicated to the study of crime in the city of São Paulo over time, with different approaches and perspectives. Among them, we highlight the sociologists Sérgio Adorno, the criminologist Guaracy Mingardi [1] [2], [3]. Some of the studies associated with the study of crime in the city of São Paulo in Brazil with computational techniques are applied by Garcia [4], [5]

These authors approach crime from different angles, such as organized crime, violence in the peripheries, the relationship between poverty and crime, the effectiveness of public security policies, among other relevant themes. His studies and research contribute to the understanding of the factors that influence crime in the city of São Paulo and to the development of more effective public policies.

In addition, authors who study data mining, such as Han and Kamber, Kononenko and Fayyad, have developed techniques for analyzing large data sets and identifying patterns and trends in different areas, including crime and [6] [7] social data [8]. These authors have proposed methodologies and tools to identify and interpret data patterns in the most varied contexts, such as clustering, decision trees and neural networks, associative patterns, among other techniques.

Thus, this study seeks to incorporate the contributions of these authors to develop a framework for the analysis of criminal data in the city of São Paulo, but not limited to it, which uses data mining techniques and approaches the complexity of this phenomenon from different perspectives. It is expected that this methodology will contribute to the development of evidence-based public policies in order to make the city of São Paulo safer and with a higher quality of life for its citizens.

Social data analytics has been widely used in a variety of fields, such as public safety, health, and finance. In the context of public safety, data analysis can help prevent and combat crime by identifying patterns and trends in criminal occurrences. In this sense, the development of a social data analysis framework can contribute to the understanding of criminal phenomena and evidence-based decision making.[9], [10]

The objective of this study is to develop a structure of analysis of social data, using criminal records of the city of São Paulo and applying descriptive computational techniques, such as associative algorithms and extraction of decision tree rules. Data analysis will be performed with a focus on discovering patterns and similarities between recurring situations in the social sphere, through descriptive techniques and data visualization. The quality of information is a central concern in the study, and a set of technological tools, such as databases, machine learning libraries, and results visualization tools, will be validated.

II. METHODOLOGY

Based on the four-step cyclical approach (acquisition, validation, projection, indication), the proposed methodology for the development of public policies to combat crime in the city of São Paulo using data mining techniques, artificial intelligence, data analysis and social techniques includes the following components demonstrated in Figure 1.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 8, August 2024

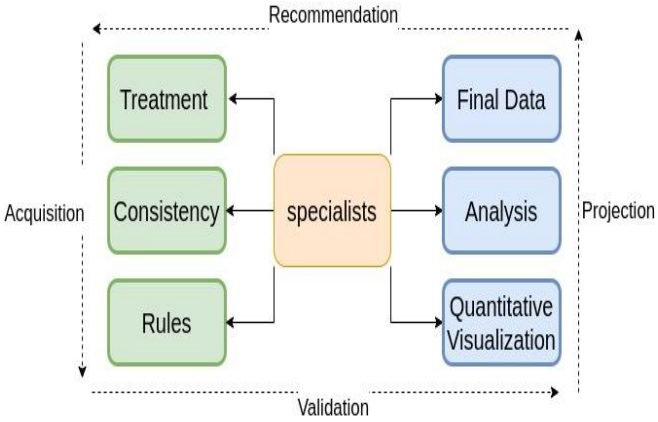

Figure 1 Framework Social Data

Components:

Treatment: includes the preparation and cleaning of the raw data, as well as the conversion to a format suitable for analysis. [11] Another point made by this component is the crossing with different sources of data related directly or indirectly to the context of evaluation. As most of the public data are anonymous in Brazil, in this component a strategy of linking by geographical location of the most varied types to be validated by specialists is adopted.

Consistency: involves validating the data to ensure that it is correct, complete, and coherent. We evaluate in this component issues related to dates, addressing and business rules relevant to the collection of data that may influence the treatment, for example, for security and respect for the laws of data processing in Brazil when the crime occurs in the residence of the victim this address is registered as null or empty. We build new attributes to contemplate this context in real data, [12] such as address attributes, which we transform into latitudes and longitudes.

Rules: define the criteria and methodologies for data analysis [13]. For each evaluation context it is necessary to establish the evaluation rules, that is, rules for defining the format that the data will be presented in the following steps, defining a technical construction from the perspective of social data specialists.

Final data: are the data ready for analysis, after undergoing treatment and consistency respecting the rules established for the evaluation context [14] .

Analysis: is the application of data mining techniques to extract relevant information from the data. [15] In this component is applied different techniques present in the areas of studies of data analysis, artificial intelligence among others, the purpose is to obtain quantitative data for a qualitative evaluation from the perspective of social data.

Quantitative visualization: involves the visual presentation of analysis results, such as graphs, tables, and maps.

Main component:

Specialists in social data: are the professionals responsible for interpreting the results of the components reported, thus, the analysis of social data related to crime. Because it is an extremely delicate subject, validation through experts is a priority in our approach even if statistically the data wants to express another result. In this way in our framework the results are not validated if the experts do not agree with some aspect generated through the set of components, thus, it is necessary, but an execution of the cycle to ensure the integrity of the partial and final deliveries of the approach.

Steps:

Acquisition: at this stage, data on crime in the city of São Paulo are collected from different sources, such as police records, demographic data, urban infrastructure data, among others. Collecting and transforming the data in order to be with integrity for a given evaluation context.

Validation: At this stage, the collected data is subjected to a validation to ensure that it is correct, complete and coherent. Data quality checks are also carried out to ensure that there are no errors or inconsistencies.

Projection: At this stage, validated data is processed and analyzed by social data experts using data mining techniques. The rules of analysis and methods for extracting relevant information from the data are defined.

Indication: In this step, the results of the analysis are presented in a quantitative visualization, such as graphs, tables and maps. Based on these results, indications and suggestions are made for the elaboration of public policies to combat crime with quantitative foundation and qualitative validation, always from the perspective of specialists in social data.

With this methodology, it is expected to obtain relevant information on crime, which can be used for the elaboration of effective public policies to combat crime and improve public safety. Making the process of discovery and use of knowledge more coherent and efficient for certain contexts of studies and evaluation.

The technology used for the development of data collection programs was Python because it has several advantages over data analysis. Python is a general programming language, which means it can be used for a variety of tasks, including data analysis, machine learning, web development, and more. Tin an active and engaged community, which means that there are many resources available to learn and utilize the language, as well as many open-source libraries for data analysis, such as NumPy, Pandas, Matplotlib, Spark, among others [15].

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 8, August 2024

III. RESULTS

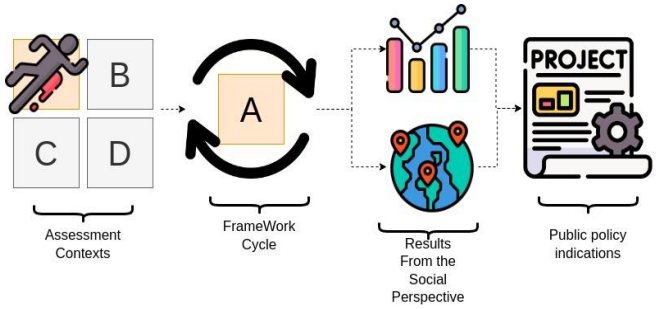

Figure 2 Frameworks Result

The results of the project to develop a structure for the analysis of social data based on criminal records of the city of São Paulo and the application of descriptive computational techniques are diverse and can bring significant contributions to the understanding of criminal phenomena and evidence-based decision making.

Some possible contributions obtained from the results of this project are:

Identification of criminal patterns and trends: The analysis of criminal data can reveal patterns and trends in relation to the most common types of crimes, places with a higher incidence of occurrences, times with a higher probability of crimes, among other relevant aspects.

Identification of risk areas: Data analysis can identify areas of risk, that is, places where the incidence of crime is higher. This information can be useful for planning public security policies.

Assistance in decision making: The results of data analysis can be used as a basis for decision making regarding the allocation of public safety resources, planning of crime-fighting operations, among other relevant aspects.

Identification of similarities between criminal situations: Data analysis can allow the identification of similar criminal situations, which can be useful for planning strategies to prevent and combat crime.

Evaluation of the effectiveness of public security policies: The results of data analysis can be used to evaluate the effectiveness of public security policies, allowing the identification of measures that have worked well and those that need to be improved.

IV. CONCLUSION

The project to develop a structure for the analysis of social data based on criminal records and the application of descriptive computational techniques presents great potential to contribute to the understanding of criminal phenomena and evidence-based decision-making in the city of São Paulo. The results obtained through the analysis of criminal data can be used for the planning and implementation of more effective and targeted public security policies. It is important to note that this project is only a starting point and that there are many possible future works, such as:

Broadening the scope of the analysis: The project can be expanded to include data from other cities or regions of the country, allowing the comparison of criminal patterns between different geographic areas.

Application of other data analysis techniques: In addition to the descriptive techniques presented in this project, other data analysis techniques, such as data mining or machine learning techniques, can be used to enhance pattern analysis and discovery.

Evaluation of the impact of public policies: In addition to evaluating the effectiveness of existing public policies, it is important to evaluate the impact of new policies implemented based on the results of data analysis.

Improving data quality: Data quality is a critical factor for the success of data analysis. It is important to ensure the quality of the data collected and stored to ensure the reliability of the results obtained.

In short, the project presented offers a set of tools and methodologies that can be explored in future work to improve the understanding of criminal phenomena and the effectiveness of public security policies.

One of the main difficulties we encountered was in relation to obtaining data, although we have laws that guarantee access to this type of data, each administrative unit in Brazil has administrative and legal autonomy, that is, each of our states and municipalities can legislate on how to disclose data if there is no national law for the subject.

Another line of thought we have reached is that a public policy of standardization of criminal information is necessary, we have a continental country and this standardization in the dissemination of crime data could contribute to the advancement and fight of crimes. However, it is necessary that this standardization occurs based on evaluations by professionals with a social perspective and not only in a numerical way.

REFERENCES

[1] C. Nonato, “Sergio Adorno: reflexões sobre a violência e a intolerância na sociedade brasileira,” *Comunicação & Educação*, vol. 20, no. 2, pp. 93–100, Oct. 2015, doi: 10.11606/ISSN.2316-9125.V20I2P93-100.

[2] R. Sérgio De Lima, S. Bueno, G. Mingardi, D. Fundação, G. Vargas, and S. Paulo -Sp -Brasil, “Estado, polícias e segurança pública no Brasil,” *Revista Direito GV*, vol. 12, no. 1, pp. 49–85, Apr. 2016, doi: 10.1590/2317-6172201603.

[3] L. E. Soares, “Segurança pública: presente e futuro,” *Estudos Avançados*, vol. 20, no. 56, pp. 91–106, 2006, doi: 10.1590/S0103-40142006000100008.

[4] G. Garcia-Zanabria and L. G. Nonato, “Visual crime pattern analysis,” *Anais Estendidos da Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 55–61, Oct. 2022, doi: 10.5753/SIBGRAPI.EST.2022.23261.

[5] G. Garcia *et al.*, “Crimanalyzer: Understanding crime patterns in São Paulo,” *IEEE Trans Vis Comput Graph*, vol. 14, no. 8, Aug. 2015, doi: 10.1109/TVCG.2019.2947515.

[6] J. Ha, M. Kambe, and J. Pe, “Data Mining: Concepts and

Techniques," *Data Mining: Concepts and Techniques*, pp. 1–703, Jan. 2011, doi: 10.1016/C2009-0-61819-5.

[7] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Mag*, vol. 17, no. 3, pp. 37–37, Mar. 1996, doi: 10.1609/AIMAG.V17I3.1230.

[8] D. de Castro Rodrigues, M. D. de Lima, and R. M. Barbosa, "Fraud detection in social income transfer programs: a social data mining approach applied to data from Brazil," *SN Social Sciences 2022 2:9*, vol. 2, no. 9, pp. 1–23, Aug. 2022, doi: 10.1007/S43545-022-00479-5.

[9] H. Hassani, X. Huang, E. S. Silva, and M. Ghodsi, "A review of data mining applications in crime," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 9, no. 3, pp. 139–154, Jun. 2016, doi: 10.1002/SAM.11312.

[10] A. A. Braga, D. Weisburd, and B. Turchan, "Focused Deterrence Strategies and Crime Control," *Criminol Public Policy*, vol. 17, no. 1, pp. 205–250, Feb. 2018, doi: 10.1111/1745-9133.12353.

[11] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, Jun. 2022, doi: 10.1016/J.GLTP.2022.04.020.

[12] S. Ackerman, "Consistency of survey opinions and external data," *Comput Stat*, vol. 34, no. 4, pp. 1489–1509, Dec. 2019, doi: 10.1007/S00180-019-00882-2/METRICS.

[13] J. Yuan, C. Chen, W. Yang, M. Liu, J. Xia, and S. Liu, "A survey of visual analytics techniques for machine learning," *Comput Vis Media (Beijing)*, vol. 7, no. 1, pp. 3–36, Mar. 2021, doi: 10.1007/S41095-020-0191-7/METRICS.

[14] M. Hosseinzadeh *et al.*, "Data cleansing mechanisms and approaches for big data analytics: a systematic study," *J Ambient Intell Humaniz Comput*, vol. 14, no. 1, pp. 99–111, Jan. 2023, doi: 10.1007/S12652-021-03590-2/METRICS.

[15] C. C. Aggarwal and J. Han, "Frequent pattern mining," *Frequent Pattern Mining*, vol. 9783319078212, pp. 1–471, Jul. 2014, doi: 10.1007/978-3-319-07821-2/COVER.