# Mapping and Predicting Crimes in Small Cities Using Web Scraping and Machine Learning

Pedro Arthur P. S. Ortiz[a] and Leandro O. Freitas[b]
*Polytechnic School, Federal University of Santa Maria, Av Roraima 1000, Santa Maria - RS, Brazil*
{*pedrops.ortiz, leanfrts*}*@gmail.com*

Keywords: Crime Prediction, Machine Learning, Web Scraping, Artificial Intelligence, Crime Analysis.

Abstract: This paper presents an approach to municipal crime analysis and prediction through the integration of web scraping techniques and artificial intelligence. Focusing on Alvorada, Brazil, we address the challenge of limited crime data availability in small cities by developing an automated system that extracts and processes crime-related information from local news sources. Our methodology employs the Anthropic Claude AI API for structured data extraction and implements a machine learning model (Random Forest) for crime prediction. The research demonstrates the feasibility of creating crime prediction systems for small cities while identifying temporal and spatial patterns in criminal activity. Additionally, we provide a framework for future improvements through potential law enforcement partnerships and dataset expansion. This study contributes to the growing field of smart city development by offering a replicable methodology for municipalities lacking standardized crime data collection systems.

## 1 INTRODUCTION

Criminality is a significant problem in cities. It has a direct impact on the quality of life of a population. Victims of violent crimes are predisposed to develop acute stress disorder and post-traumatic stress disorder, which are frequently associated with depression and anxiety (Lefebvre et al., 2021). Tourism is the biggest service industry and a fundamental part of numerous economies. Crime and tourism share a complex two-way relationship that not only impacts the area's economic growth but also affects society and individuals directly. This interplay demands careful attention for both state policy-making and the organization of law enforcement agencies (Shchokin et al., 2023). The concept of smart cities emerges as a potential solution, leveraging information and communication technologies to create favorable living conditions for residents and visitors while promoting economic activity and security. Recent studies in Polish provincial cities demonstrate how comprehensive security analysis can inform urban safety strategies through various crime indicators and multi-criteria analysis methods (Tutak and Brodny, 2023). Despite its importance, tracking and analyzing urban crime presents significant challenges, particularly at the municipal level. However, artificial intelligence has shown promising results in urban analysis and decision-making, as demonstrated by studies in Morocco where AI models achieved over 80% accuracy in predicting urban attractiveness factors (Khalid et al., 2023).

Despite the potential of smart city solutions for security, several challenges remain before widespread implementation becomes feasible. One significant obstacle is the lack of standardised city-level data on criminal activity (Muggah and Aguirre, 2024). Traditional criminological data collection methods are often too slow and resource-intensive to provide timely insights for decision-makers. Web scraping emerges as a promising complementary tool to address this data scarcity, particularly at the municipal level. When combined with machine learning models, scraped data can be used to train predictive systems that analyze crime patterns and generate probability statistics. This approach can serve as a valuable auxiliary tool for crime analysts and urban planners, supporting their efforts to develop safer smart cities through data-driven decision making.

This paper presents an exploratory study into the potential of web scraping and machine learning techniques for crime mapping and prediction. We focus on Alvorada, Rio Grande do Sul, Brazil, as our case

[a] https://orcid.org/0009-0002-2522-9568
[b] https://orcid.org/0000-0002-1112-3685



Ortiz, P. A. P. S. and Freitas, L. O.
Mapping and Predicting Crimes in Small Cities Using Web Scraping and Machine Learning.
DOI: 10.5220/0013421400003929

In *Proceedings of the 27th International Conference on Enterprise Information Systems (ICEIS 2025) - Volume 1*, pages 888-894
ISBN: 978-989-758-749-8; ISSN: 2184-4992

study. This choice is supported by two key factors: first, as a small city, it lacks standardized crime data, making it an ideal candidate for testing alternative data collection methods. Second, its significant historical context - having once been ranked as Brazil's sixth most violent city (Mendes, 2024), before achieving its lowest homicide rates in recent history - provides a rich background for analysis. Although our study centers on Alvorada, the framework we have developed can be applied to any city lacking standardized crime data. Our approach encompasses the following steps:

1. Data Collection: Web scraping of crime-related news from two local news platforms (G1[1] and O Alvoradense[2]) to create a comprehensive crime incident dataset.
2. AI-Powered Data Analysis: Implementation of advanced AI techniques to extract and process relevant information from the collected news articles, converting unstructured text data into a structured dataset with standardized crime categories and details.
3. Geographical Mapping and Analysis: Generation of crime heat maps and spatial analysis tools to identify potential hotspots and patterns across the city, providing visual insights into crime distribution.
4. Predictive Modeling: Development and application of machine learning models to analyze crime patterns and generate probability-based predictions, offering insights into potential future crime trends and risk areas.

Through the application of this framework, we address two key challenges: the lack of standardized crime data at the municipal level and the potential integration of web scraping and machine learning tools into smart city safety planning. While our approach provides innovative solutions for data collection and analysis, it is designed to augment rather than replace human analysts, serving as a complementary tool in their decision-making process. Our methodology focuses specifically on crime pattern analysis and prediction, acknowledging that the broader socioeconomic causes of criminal behavior lie beyond the scope of this research.

This paper is organized as follows: Section 2 presents related work in crime prediction and pattern analysis; Section 3 discusses the state of the art in web scraping and AI applications for crime analysis; Section 4 details our methodology and experimental setup, including the implementation of web scraping and machine learning components; Section 5 presents our results and analysis of the crime dataset; and finally, Section 6 concludes the paper and discusses future work.

[1]https://g1.globo.com/

[2]https://oalvoradense.com.br/

## 2 RELATED WORK

A comprehensive scientometric analysis of crime prediction and pattern analysis (CPPA) research over the past decade reveals the growing integration of artificial intelligence in this field. According to the analysis, researchers have developed five distinct clusters of AI methodologies applied to crime prediction, demonstrating the field's rapid evolution and diversification. The study highlights how the combination of AI with traditional criminological approaches has enabled more sophisticated and accurate crime prediction models (Kaur and Saini, 2024).

The analysis particularly emphasizes the emergence of several key research trends: the increasing use of machine learning algorithms for pattern recognition in criminal behavior, the development of predictive modeling techniques, and the growing importance of data preprocessing and feature selection in crime analysis. These findings align with our approach of combining web scraping with AI analysis, particularly in addressing the challenges of data collection and standardization in small cities.

What makes this analysis particularly relevant to our work is its identification of the growing trend toward integrating diverse data sources and AI methodologies in crime prediction. While many studies focus on large urban centers with established data collection systems, our work extends these principles to small cities where traditional data sources may be limited or unavailable. This adaptation of advanced AI techniques to local contexts represents an important evolution in the field of crime prediction and analysis.

## 3 STATE OF THE ART

Web scraping typically involves three basic steps - fetching, extracting, and transforming data - but it is often treated as a peripheral tool rather than a core methodology in research (NR et al., 2023). Our study takes a more comprehensive approach, particularly given the sensitive nature of crime data collection and analysis. We propose an enhanced web scraping framework (illustrated in Figure 1) that emphasizes careful supervision and methodological rigor throughout the data collection process.

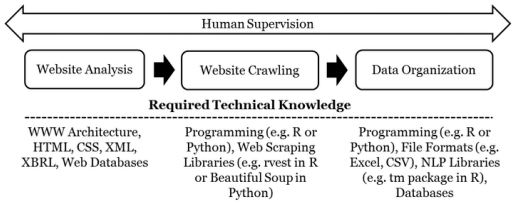


Figure 1: The Web Scraping Process (Krotov and Tennyson, 2018).

The Anthropic Claude AI API represents a significant advancement in natural language processing and structured data extraction. As a large language model, it excels at understanding context and extracting relevant information from unstructured text, making it particularly suitable for processing news articles and converting them into structured crime data. The API's capabilities include entity recognition, relationship extraction, and contextual understanding, which are essential for accurate crime data categorization.

Artificial intelligence serves as a valuable tool in urban analysis and decision-making (Khalid et al., 2023). AI models can achieve significant accuracy in analyzing urban factors and making predictions about city development patterns. In the context of criminology, artificial intelligence has shown particular promise in crime prediction and pattern analysis (CPPA), with research identifying five distinct clusters of AI methodologies being applied to this field (Kaur and Saini, 2024).

The current applications of AI in crime analysis focus on several key areas (Gül, 2024):

- Pattern Recognition: AI systems can analyze crime data to identify recurring patterns and trends
- Predictive Modeling: Modern algorithms can process historical data to generate predictions about potential crime occurrences
- Data Processing: AI assists in standardizing and analyzing large volumes of crime-related data from various sources

These capabilities are particularly relevant for our study, as they enable the processing and analysis of web-scraped crime data from news sources, helping to address the challenge of standardized data scarcity in small cities.

# 4 METHODOLOGY AND EXPERIMENTAL SETUP

Our primary objectives are to extract data from two news media sources (G1 and O Alvoradense) for comprehensive data collection, to clean and process this news data using the Anthropic Claude AI API to create a structured criminological dataset, and to utilize this dataset for both mapping and potential crime prediction in the city of Alvorada. According to the 2022 census, Alvorada has an area of 70.811 km² and a population of 187,315 inhabitants (IBGE, 2024). Our selection of Alvorada as a case study is supported by two key factors: the absence of standardized crime analysis data and its historical significance as Brazil's sixth most violent city (Mendes, 2024).

## 4.1 Web Scraping Implementation

The implementation of our web scraping framework follows a modular approach, utilizing Python with Selenium WebDriver for dynamic content extraction. The system architecture comprises several key components, each serving a specific function in the data collection process.

### 4.1.1 System Architecture

The scraping system is built with the following components:

- Browser Automation: Selenium WebDriver with Chrome in headless mode
- Data Storage: CSV file integration with Google Drive for persistent storage
- Rate Limiting: Implemented through strategic delays to prevent server overload

The collected data is structured in a standardized format with the following fields:

- title: Article headline
- link: URL to the full article
- source: Source identification
- data_collection: Timestamp of data collection

### 4.1.2 Data Processing Cleaning Methodology

Let $N$ be the set of all collected news articles, where each article $n_i \in N$ is characterized by a tuple:

$$n_i = (t_i, l_i, s_i, e_i) \quad (1)$$

where $t_i$ represents the title, $l_i$ the link, $s_i$ the source, and $e_i$ the extraction timestamp.

The classification of crime-related content is defined through two primary sets:

1. Pattern Set $P$: A collection of regular expressions $p_j$ covering different crime categories:

$$P = \{p_1, p_2, ..., p_m\} \text{ where } m = 7 \text{ categories} \quad (2)$$

2. Keyword Set $K$: A finite set of crime-related terms:

$$K = \{k_1, k_2, ..., k_r\} \text{ where } r = |K| = 40 \text{ terms} \quad (3)$$

The crime classification function $C(t_i)$ for any title $t_i$ is defined as:

$$C(t_i) = \begin{cases} 1 & \text{if } (\exists p_j \in P : p_j \text{ matches } t_i) \vee (W(t_i) \cap K \neq \emptyset) \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where $W(t_i)$ is the set of words in title $t_i$ after normalization.

The resulting filtered dataset $D_c$ contains all crime-related articles:

$$D_c = \{n_i \in N : C(t_i) = 1\} \quad (5)$$

#### 4.1.3 AI-Powered Data Extraction

Given our filtered crime dataset $D_c$, we define an AI-powered transformation function $\mathcal{T}$ that maps each news article to a structured crime record:

$$\mathcal{T} : D_c \to S \quad (6)$$

where $S$ is the space of structured crime records. Each structured record $s_i \in S$ is defined as a 7-tuple:

$$s_i = (u_i, d_i, l_i, c_i, t_i, w_i, a_i) \quad (7)$$

where: - $u_i \in [1000, 9999]$ : unique identifier - $d_i \in \mathcal{D} \cup \{N/A\}$ : datetime of incident - $l_i \in \mathcal{L} \cup \{N/A\}$ : location hierarchy - $c_i \in \mathcal{C}$ : city identifier - $t_i \in \mathcal{P}(\text{CRIME\_TYPES})$ : set of crime types - $w_i \in \mathcal{W} \cup \{N/A\}$ : weapon used - $a_i \in \{True, False, Unknown\}$ : arrest status

The location hierarchy $\mathcal{L}$ is defined as an ordered set:

$$\mathcal{L} = \{Street, Neighborhood, Region\} \quad (8)$$

with precedence relation:

$$Street \succ Neighborhood \succ Region \quad (9)$$

The transformation process $\mathcal{T}$ involves content extraction $\xi$ and AI analysis $\alpha$:

$$\mathcal{T}(n_i) = \alpha(\xi(n_i)) \quad (10)$$

Content length constraint:

$$|\xi(n_i)| \leq 4000 \text{ characters} \quad (11)$$

The final structured dataset $D_s$ is defined as:

$$D_s = \{\mathcal{T}(n_i) : n_i \in D_c\} \quad (12)$$

#### 4.1.4 Feature Engineering

The feature space $\mathcal{F}$ consists of both temporal and contextual variables:

$$\mathcal{F} = \{f_t\} \cup \{f_c\} \quad (13)$$

where $f_t$ represents temporal features and $f_c$ represents contextual features:

Temporal Features $f_t$:

- $T_{day} \in \{0,1,2,3,4,5,6\}$ : day of week
- $T_{month} \in \{1, ..., 12\}$ : month
- $T_{hour} \in \{-1, 0, ..., 23\}$ : hour of day
- $T_{season} \in \{0,1,2,3\}$ : encoded season

Contextual Features $f_c$:

- $B_{weekend}, B_{holiday}, B_{night}, B_{rush} \in \{0,1\}$ : binary indicators
- $B_{start\_month}, B_{end\_month} \in \{0,1\}$ : monthly period indicators
- $B_{business}, B_{lunch} \in \{0,1\}$ : activity period indicators

#### 4.1.5 Model Architecture

Our Random Forest model implementation follows an ensemble learning approach with carefully tuned hyperparameters for optimal performance in crime prediction:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} f_b(x) \quad (14)$$

The model architecture includes:

- Base Estimator: Decision Trees with Information Gain criterion
- Number of Estimators: 200 trees in the forest
- Maximum Depth: 30 levels per tree
- Minimum Samples Split: 5 samples required to split internal node
- Feature Selection: Square root of total features considered at each split
- Bootstrap Sampling: Enabled for training individual trees

This configuration was chosen based on extensive cross-validation testing and optimization for our specific crime prediction task. The relatively high number of estimators (200) helps ensure robust predictions while avoiding over fitting, while the maximum depth of 30 allows the model to capture complex patterns in the crime data.

# 5 RESULTS AND CRIME DATASET ANALYSIS

Our analysis revealed striking patterns in the spatial distribution of criminal activities across Alvorada. The heatmap visualization uncovered distinct vulnerability zones, with particularly intense clustering in the eastern and central regions of the city. The most prominent concentration appears as a deep red zone in the central-eastern area, suggesting this location experiences significantly higher crime rates than surrounding neighborhoods. A secondary cluster of moderate intensity (shown in yellow) emerges slightly north of the primary area, while several lower-intensity areas (depicted in blue) radiate outward from these central points.

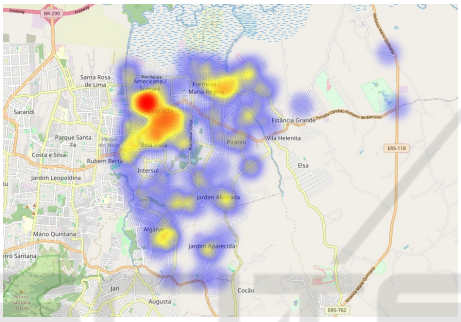

Figure 2: Alvorada Crime Heatmap.

As shown in Figure 2, this spatial pattern offers valuable insights for both public safety planning and urban development strategies. The identification of these higher-activity zones enables more effective allocation of public resources and community support programs. The concentration of activities appears to follow specific geographic patterns that correlate with commercial density, public transportation routes, socioeconomic variations across neighborhoods, and overall population density. Understanding these patterns creates opportunities for collaborative solutions between law enforcement, urban planners, and community leaders, potentially addressing root causes rather than symptoms.

Moving beyond spatial analysis, our temporal investigation uncovered sophisticated patterns across different time scales. Our examination revealed distinct behavioral differences between property crimes and violent crimes throughout the day.

As illustrated in Figure 3, property crimes show a pronounced peak during morning hours, particularly between 8:00 and 10:00, reaching probability rates of up to 90%. This morning surge in property crimes likely coincides with times when residences are most vulnerable as occupants leave for work or school. In

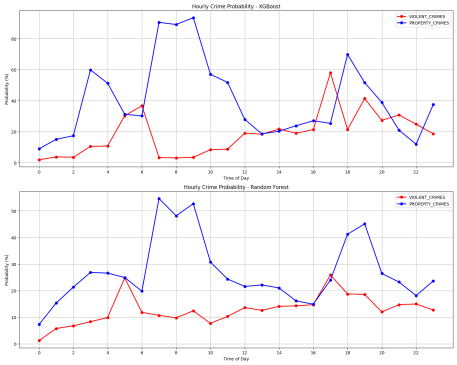


Figure 3: Hourly Crime Distribution.

contrast, violent crimes exhibit a markedly different temporal signature, with significant spikes during the evening hours, particularly around 17:00-18:00, with probability rates reaching nearly 60%. This evening peak in violent crimes could be associated with increased social interaction during after-work hours and nighttime activities.

Expanding our temporal analysis to a broader scale reveals equally significant patterns throughout the year.

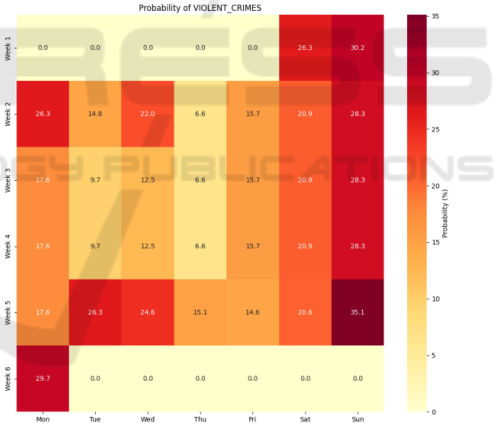


Figure 4: Monthly Violent Crimes Possibility.

The monthly patterns of crime, depicted in Figure 4, demonstrate notable seasonal variations in criminal activity. During warmer months, particularly between October and March, the data indicates elevated incident rates across multiple crime categories. Further examination of crime probability evolution throughout each month reveals intricate patterns in criminal behavior, with property crimes consistently showing the highest probability rates, especially during the final days of the month.

The distribution of different crime categories, presented in Figure 5, reveals significant variations in crime type frequencies across months. Property

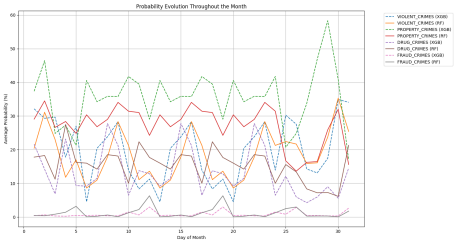


Figure 5: Distribution of Crime Types by Month.

crimes and violent crimes display distinct seasonal patterns, with property crimes showing higher frequencies during summer months and violent crimes exhibiting more uniform distribution throughout the year. Drug-related offenses show moderate but consistent presence, while fraud-related crimes maintain lower but stable frequencies across all periods.

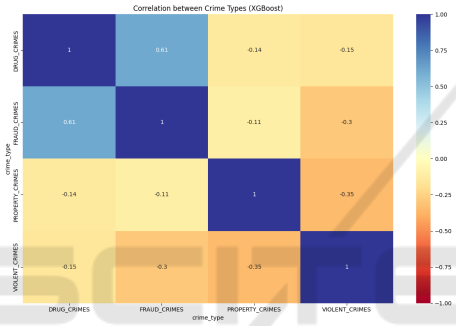


Figure 6: Correlation Matrix of Crime Factors.

To validate and quantify these observed patterns, we analyzed the correlations between various crime factors. The correlation matrix presented in Figure 6 reveals significant relationships between temporal, spatial, and environmental variables. This comprehensive analysis demonstrates how various urban factors interconnect to influence criminal activity patterns, providing statistical validation for our observations and offering valuable insights for developing more nuanced and effective public safety strategies.

Our predictive models, while showing moderate performance with accuracy rates of 0.25-0.26, demonstrate the potential of this approach for crime prediction in small cities. The consistent patterns identified by Random Forest model, despite their different algorithmic approaches, lend credibility to these findings and suggest that meaningful patterns in criminal activity can be captured and predicted, though there remains significant room for improvement through enhanced data collection and feature engineering.

# 6 CONCLUSIONS AND FUTURE WORK

This research demonstrates the feasibility of developing crime prediction models for small cities, with particularly noteworthy results in temporal and spatial pattern identification. Our implementation not only established an automated data collection and processing pipeline but also revealed distinct crime patterns that could significantly impact public safety strategies. The spatial analysis identified clear vulnerability zones in Alvorada's eastern and central regions, with the heatmap visualization revealing concentrated areas of criminal activity that correlate strongly with urban infrastructure patterns.

The temporal analysis yielded several crucial insights, particularly in differentiating between property and violent crimes. The discovery that property crimes peak during morning hours (8:00-10:00) with up to 90

Our methodological approach, combining web scraping with AI-powered analysis, proved particularly effective in extracting structured information from unstructured news sources. The Random Forest model, while showing moderate performance with accuracy rates of 0.25-0.26, demonstrated consistent pattern recognition capabilities across different algorithmic approaches. This consistency lends credibility to the identified patterns and suggests that the methodology is sound, even if there is room for improvement in prediction accuracy.

The current implementation faces several challenges that warrant attention. The limited dataset size of approximately 750 valid records impacts the model's predictive accuracy, suggesting that expanding the dataset would likely improve performance significantly. The correlation analysis between crime factors revealed complex interconnections between temporal, spatial, and environmental variables, indicating that a larger dataset would allow for more nuanced pattern detection and improved prediction accuracy.

Future work should focus on three key areas:

1. Data Enhancement: - Establish formal partnerships with law enforcement agencies to expand the dataset to 3,000-5,000 records - Integrate official police reports and statistics to complement news-based data - Implement real-time data collection mechanisms for continuous model updating

2. Methodological Improvements: - Develop more sophisticated feature engineering techniques based on the identified correlations - Enhance spatial analysis methods to better capture neighborhood-specific patterns - Integrate deep learning approaches

for improved pattern recognition - Refine natural language processing capabilities for better information extraction

3. Operational Integration: - Develop interactive visualization tools for law enforcement use - Create real-time prediction capabilities for immediate response - Establish standardized reporting protocols for consistent data collection - Implement robust quality control measures for data validation

The establishment of data-sharing protocols with law enforcement would not only enhance data quality but also enable the validation of predictions against actual crime reports. This feedback loop would be crucial for continuous system improvement and adaptation to evolving crime patterns.

Our findings suggest that while predictive accuracy can be improved, the current system already provides valuable insights for urban safety planning. The clear temporal and spatial patterns identified could immediately inform resource allocation decisions, even as the system continues to evolve. The correlation analysis particularly highlights the interconnected nature of urban factors in crime occurrence, suggesting that a holistic approach to crime prevention, involving both law enforcement and urban planning strategies, would be most effective.

This research lays the groundwork for more sophisticated crime prediction systems in small cities, particularly those with limited resources for dedicated crime analysis infrastructure. The proposed developments, especially the expansion of the dataset through official partnerships, could transform this prototype into a powerful tool for law enforcement and urban planning applications. As cities continue to evolve and face new security challenges, such data-driven approaches will become increasingly valuable for maintaining public safety and optimizing resource allocation. The framework developed here offers a replicable methodology that other small municipalities can adapt and implement, contributing to the broader goal of creating safer, more resilient urban environments through data-driven decision making.

## REFERENCES


Gül, R. (2024). What is state-of-art artificial intelligence? Accessed: 2024-12-10.

IBGE (2024). Alvorada. Retrieved December 11, 2024.

Kaur, M. and Saini, M. (2024). Role of artificial intelligence in the crime prediction and pattern analysis studies published over the last decade: a scientometric analysis. *Artificial Intelligence Review*, 57(8):202.

Khalid, S., Effina, D., Khalid, K. R., and Chaabane, M. S. (2023). The artificial intelligence as a decision-making instrument for modeling and predicting small cities' attractiveness: evidence from morocco. *Appl. Math. Inf. Sci*, 17(5):905–913.

Krotov, V. and Tennyson, M. (2018). Scraping financial data from the web using the r language. *Journal of Emerging Technologies in Accounting*. Forthcoming.

Lefebvre, C., Fortin, C., and Guay, S. (2021). Quality of life after violent crime: the impact of acute stress disorder, posttraumatic stress disorder, and other consequences. *Journal of Traumatic Stress*, 34(3):526–537.

Mendes, L. (2024). Alvorada, que já foi a sexta cidade mais violenta do país, chega ao menor número de homicídios da história. Accessed: 2024-12-10.

Muggah, R. and Aguirre, K. (2024). Latin america's murder rates reveal surprising new trends. *Americas Quarterly*. Accessed: [Insert date of access here].

NR, R. R., Vijayalakshmi, M., et al. (2023). Web scrapping tools and techniques: A brief survey. In *2023 4th International Conference on Innovative Trends in Information Technology (ICITIIT)*, pages 1–4. IEEE.

Shchokin, R., Maziychuk, V., Mykhailyk, O., Kolomiiets, A., Akifzade, S., and Tymoshenko, Y. (2023). The impact of the crime rate on the hospitality and tourism industry in the eu countries. *Geo Journal of Tourism and Geosites*, 46(1):135–147.

Tutak, M. and Brodny, J. (2023). A smart city is a safe city: analysis and evaluation of the state of crime and safety in polish cities. *Smart Cities*, 6(6):3359–3392.