Extração de Notícias sobre Segurança Pública para Desenvolvimento de Corpora em português: uma análise preliminar em cidades do nordeste brasileiro

***Matheus Ryan da Silva Nascimento[1], Vagner Alves Ferreira da Silva[1], Gabriel Souza[1], Kauã Gabriel Silva de Lima[1], Ericlécio Thiago Morais de Araújo[1], Everton Reis de Souza[1], Jean Turet[1] e Victor Diogho Heuer de Carvalho[1]**

[1]Group of Engineering in Decision-Making and Artificial Intelligence (GEDAI-UFAL) - Campus do Sertão, Universidade Federal de Alagoas. Rodovia AL 145, Km 3, nº 3849, Cidade Universitária. Delmiro Gouveia, Alagoas, Brasil CEP: 57480-000

Resumo: *Esta pesquisa concentra-se na coleta de artigos de notícias relacionados à segurança pública para a construção de um corpus abrangente em português. Atualmente, o estudo está na fase de aquisição e processamento de textos noticiosos por meio de web scraping em sites e blogs, trazendo uma análise preliminar sobre os dados levando em consideração algumas cidades Brasileiras, adotando como metodologia a compreensão e estrutura dos sites, definição dos termos e buscas, armazenamento, processamento e análise dos dados. O principal objetivo é criar um recurso linguístico que possa ser utilizado em diversas aplicações de processamento de linguagem natural (PLN) no futuro. O corpus resultante servirá de base para o desenvolvimento de ferramentas e tecnologias capazes de analisar e compreender temas relacionados à segurança pública na língua portuguesa, contribuindo para avanços na área e para uma melhor compreensão desse cenário.*

Abstract. *This study aims to collect news articles related to public security in order to build a comprehensive Portuguese-language corpus. The research is currently in the data acquisition and preprocessing phase, which involves web scraping news content from websites and blogs. A preliminary analysis of the collected data is being conducted, focusing on selected Brazilian cities. The adopted methodology includes understanding and mapping the structure of target websites, defining relevant search terms, and implementing strategies for data storage, processing, and analysis. The primary objective is to develop a linguistic resource that can support a wide range of natural language processing (NLP) applications. The resulting corpus will serve as a foundation for the creation of tools and technologies capable of analyzing and interpreting public security-related content in Portuguese. This initiative is expected to contribute to technological advancements in the field and to a deeper understanding of public security discourse in Brazil.*

1. Introdução

A segurança pública continua sendo uma questão crítica nas sociedades contemporâneas, com o aumento das taxas de criminalidade e violência representando desafios

significativos para governos e comunidades. Garantir a segurança e o bem-estar dos cidadãos é uma preocupação primordial que exige estratégias e políticas eficazes. No entanto, a complexidade e a natureza multifacetada dos problemas de segurança pública exigem uma abordagem abrangente que vá além das fontes de dados tradicionais [Turet & Costa, 2022].

Nesse contexto, dados não estruturados, como artigos de notícias, postagens em redes sociais e entradas de blogs, desempenham um papel fundamental ao fornecer insights valiosos para análise e tomada de decisão. Diferentemente dos dados estruturados, os dados não estruturados oferecem uma perspectiva rica e diversificada sobre questões de segurança pública, capturando eventos em tempo real, sentimentos e tendências frequentemente negligenciadas. Ao aproveitar essas fontes de dados, pesquisadores e formuladores de políticas podem desenvolver estratégias mais informadas e responsivas para enfrentar os desafios da segurança pública [Turet & Costa, 2022].

Vários estudos destacaram a importância do uso de dados não estruturados na segurança pública. Por exemplo, [Suhaimin et.al, 2022] examinou a análise de sentimento em redes sociais e mineração de opiniões, identificando tendências e questões relacionadas à segurança pública. Sua pesquisa enfatizou o potencial da análise de mídias sociais para aprimorar medidas de segurança pública. Além disso, [Carnaz et.al, 2021] criou um corpus anotado de documentos em português relacionados a crimes, facilitando aplicações de processamento de linguagem natural (PLN) e aprendizado de máquina na segurança pública. Ademais, [Turet & Costa, 2022] desenvolveu uma metodologia híbrida para analisar dados estruturados e não estruturados a fim de apoiar a tomada de decisões na segurança pública, demonstrando um aumento de 80% na precisão dos algoritmos de previsão de crimes. [Silva et.al, 2021] realizou uma análise de sentimento de dados de redes sociais para previsão de crimes no Brasil, demonstrando a eficácia das técnicas de aprendizado de máquina na identificação de atividades criminosas em potencial. [Gomes et.al, 2021] explorou o uso de aprendizado de máquina para detectar cyberbullying em redes sociais, destacando a importância de sistemas automatizados na identificação de comportamentos prejudiciais online. Esses exemplos ilustram o potencial dos dados não estruturados para aprimorar nossa compreensão da dinâmica da segurança pública.

Além disso, a importância da construção de corpora em português para a análise da segurança pública tem sido enfatizada em pesquisas recentes. [de Carvalho et.al, 2022; de Carvalho et.al, 2022] apresentaram dois corpora do português brasileiro coletados de diferentes mídias sobre questões de segurança pública em uma localização específica. Seu estudo destacou o valor desses corpora no suporte a análises e processos de tomada de decisão para as autoridades de segurança. [de Carvalho et.al, 2023] fornece uma revisão abrangente das aplicações de mineração de texto e análise na segurança pública, identificando principais áreas de aplicação, ferramentas tecnológicas recorrentes e direções para pesquisas futuras.

Com base nisso, este estudo busca contribuir para essa iniciativa por meio da coleta e processamento de textos de notícias relacionados à segurança pública em português, utilizando *web scraping* de vários sites e blogs. O objetivo principal é construir um corpus abrangente que possa servir como um recurso linguístico valioso para aplicações de

processamento de linguagem natural (PLN). Esse corpus apoiará o desenvolvimento de ferramentas e tecnologias capazes de analisar e compreender tópicos de segurança pública, aprimorando, assim, a capacidade de responder de maneira eficaz às questões de segurança pública em regiões brasileiras.

2. Materiais e Métodos

Dada a natureza crítica da segurança pública nas sociedades contemporâneas, este estudo busca aproveitar fontes de dados não estruturados para fornecer insights valiosos para processos de análise e tomada de decisão. A pesquisa envolveu a seleção de 20 sites e blogs focados em segurança pública, divididos entre quatro cidades do Brasil: Maceió, Recife, Caruaru e Natal considerando o período de dois anos. As cidades de Maceió, Recife, Caruaru e Natal foram selecionadas por representarem diferentes contextos urbanos do Nordeste brasileiro, combinando capitais e cidade do interior. A escolha considerou a alta incidência de violência, a diversidade social e a disponibilidade de fontes locais de notícias, permitindo uma análise comparativa e aprofundada sobre segurança pública com base em dados não estruturados. O objetivo principal foi identificar notícias relacionadas à violência em um contexto geral. A metodologia adotada abrangeu várias etapas, conforme detalhado a seguir [de Carvalho et.al, 2022; Turet & Costa, 2023]:

(i) Seleção de Fontes: Foram escolhidos 20 sites e blogs relevantes que discutem questões de segurança pública nas quatro cidades selecionadas. Os critérios de seleção incluíram a frequência de atualizações, a relevância das informações e o alcance do público. Garantir uma ampla diversidade de fontes ajudou a capturar uma perspectiva abrangente sobre as questões de segurança pública.

(ii) Compreensão da Estrutura dos Sites: A equipe de pesquisa analisou minuciosamente a estrutura de cada site e blog para entender como as informações são organizadas e publicadas. Essa análise incluiu a identificação de seções específicas, categorias de notícias e padrões de formatação. Compreender a estrutura foi essencial para o desenvolvimento de um algoritmo eficiente de web scraping.

(iii) Definição de Termos de Busca: Foram definidos termos de busca específicos relacionados à violência, como "crime", "assalto", "homicídio" e "tráfico de drogas". Esses termos foram utilizados para filtrar artigos de notícias relevantes. Ao focar em palavras-chave precisas, a pesquisa buscou capturar um conjunto abrangente de dados que refletisse eventos em tempo real, sentimentos e tendências frequentemente negligenciadas por fontes de dados estruturadas.

(iv) Desenvolvimento do Algoritmo de Web Scraping: Um algoritmo sofisticado de web scraping foi desenvolvido para coletar automaticamente notícias dos sites e blogs selecionados. O algoritmo foi projetado para navegar pelas páginas da web, extrair conteúdos relevantes e armazenar os dados em um formato estruturado. Esse processo automatizado garantiu a coleta eficiente e precisa de grandes volumes de dados.

(v) Armazenamento dos Dados: Os dados coletados foram armazenados em arquivos JSON, garantindo que todas as informações estivessem organizadas e acessíveis para

futuras análises. A escolha do formato JSON facilitou a manipulação e o processamento dos dados.

(vi) Processamento dos Dados: Dado que os dados coletados eram não estruturados, foi necessário um processamento preliminar. Isso incluiu a limpeza dos dados, remoção de duplicatas, correção de erros e padronização das informações para facilitar a análise. O objetivo do processamento foi melhorar a qualidade dos dados e garantir a consistência do conjunto de informações.

(vii) Análise e Visualização dos Dados: Após o processamento, os dados foram analisados e visualizados por meio de gráficos e outras representações visuais. Essas visualizações ajudaram a identificar padrões, tendências e insights sobre a violência nas quatro cidades. Ao utilizar dados não estruturados, a pesquisa buscou desenvolver estratégias mais informadas e responsivas para enfrentar os desafios da segurança pública de maneira eficaz.

Essas etapas destacam o potencial dos dados não estruturados para aprimorar a compreensão da dinâmica da segurança pública e apoiar processos de tomada de decisão mais eficazes.

3. Resultados e discussões iniciais

Como resultado dessa abordagem abrangente, aproximadamente 65.000 artigos de notícias relacionados à violência foram identificados e coletados. Esse corpus fornece um conjunto de dados diversificado para a análise de questões de segurança pública, capturando eventos em tempo real, sentimentos e tendências em quatro cidades brasileira.

Abaixo encontra-se um exemplo das nóticias que foram capturadas:

{"data":"sábado, 27 de maio de 2023","conteudo":"Preso integrava quadrilha e tem condenação por crime de homicídio e tráfico de drogas em Recife A Polícia Federal em Pernambuco, através de sua representação regional da Interpol, promoveu a extradição de um fugitivo brasileiro, condenado no Brasil a 20 anos de prisão por homicídio e tráfico de drogas. O suspeito é nadador, natural de Recife\/PE e tem 28 anos. O voo trazendo o recifense chegou ontem (25\/05) por volta das 22h no aeroporto Internacional"}

A **Figura 1** a seguir ilustra a distribuição dos artigos de notícias coletados entre as cidades de **Recife, Maceió, Natal e Caruaru**.

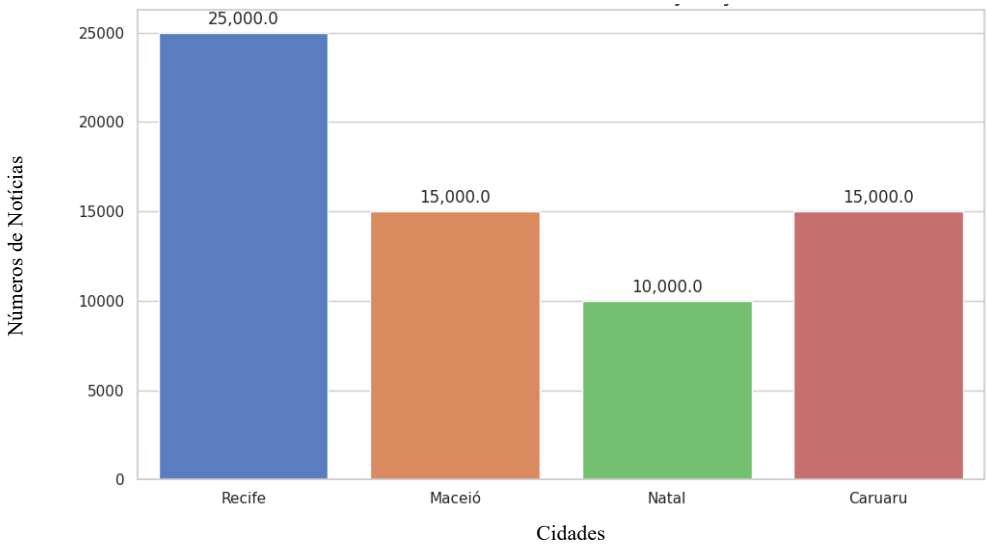

Figura 1. Número de Artigos de Notícias por Cidade

Inicialmente, para ilustrar os temas mais frequentes nas notícias coletadas, construímos uma nuvem de palavras que destaca os termos relacionados à violência e segurança pública presentes no corpus de aproximadamente 65.000 artigos. Essa visualização sintetiza as palavras mais recorrentes, oferecendo uma percepção imediata dos tópicos mais relevantes discutidos pela mídia nas cidades analisadas (Figura 1).

Figura 1. Nuvem de Palavras

A nuvem de palavras acima representa visualmente os principais termos encontrados nas aproximadamente 65.000 notícias relacionadas à violência coletadas nas cidades de Recife, Maceió, Natal e Caruaru. Palavras como **crime**, **assalto**, **homicídio** e **tráfico de drogas** aparecem com maior destaque, refletindo a frequência e relevância desses temas no corpus analisado.

Essa representação gráfica facilita a compreensão das temáticas mais recorrentes nas notícias, evidenciando os desafios centrais enfrentados na segurança pública dessas regiões. Termos relacionados a estratégias de intervenção, como **prevenção**, **comunidade** e **polícia**, também são visíveis, indicando a preocupação com medidas para combater a violência e proteger a população.

Especificamente quanto aas notícias capturadas, em **Recife**, foram coletados **25.000** artigos de notícias. Sendo uma das maiores cidades da região, Recife naturalmente possui um volume mais alto de cobertura jornalística. O tamanho da cidade, a densidade populacional e as diversas condições socioeconômicas contribuem para uma maior incidência de eventos violentos reportados, proporcionando uma visão abrangente dos desafios de segurança pública enfrentados pela cidade.

Em **Maceió**, foram coletados **15.000** artigos de notícias. Os desafios de segurança pública da cidade, incluindo taxas de criminalidade e instabilidade social, são frequentemente abordados pela mídia local, refletindo as preocupações contínuas de seus moradores.

Em **Natal**, foram coletados **10.000** artigos de notícias. Embora menor em comparação com Recife e Maceió, Natal ainda gera uma quantidade significativa de notícias relacionadas à violência. A importância estratégica da cidade e seu papel como um centro regional contribuem para sua cobertura midiática, embora em menor escala do que as cidades maiores.

Em **Caruaru**, foram coletados **15.000** artigos de notícias. Apesar de não ser tão grande quanto Recife ou Maceió, Caruaru apresenta um número de artigos semelhante ao de Maceió. Isso indica que a segurança pública é uma preocupação relevante na cidade, com a mídia local cobrindo ativamente diversos incidentes e questões relacionadas à violência.

Para obter uma compreensão mais aprofundada dos tipos de violência reportados nos artigos coletados, analisamos a distribuição de termos específicos relacionados à violência, incluindo **"crime"**, **"assalto"**, **"homicídio"** e **"tráfico de drogas"**.

O termo **"crime"** é o mais frequente, com **30.000 ocorrências**, abrangendo diversos atos violentos, incluindo feminicídio. Esse escopo amplo explica sua alta frequência, destacando a complexidade dos desafios de segurança pública. O termo **"assalto"** aparece **15.000 vezes**, indicando preocupações significativas com a violência física. **"Homicídio"**, com **10.000 ocorrências**, enfatiza a gravidade dos crimes violentos fatais. **"Tráfico de drogas"** também registra **10.000 ocorrências**, evidenciando a persistência da violência relacionada ao tráfico de entorpecentes.

Especificamente com relação a prevalência do termo **"crime",** este reflete a natureza multifacetada dos desafios da segurança pública, pois engloba uma ampla gama de atividades criminosas. Essa categorização ampla é essencial para capturar todo o espectro de atos violentos que impactam a segurança da população.

De forma geral, essa análise da distribuição dos termos fornece uma visão inicial sobre os diferentes tipos de violência que afetam as quatro cidades estudadas. Essas informações são fundamentais para que pesquisadores e formuladores de políticas desenvolvam e implementem estratégias de segurança pública eficazes (em alinhamento

com outros dados e estudos), adaptadas às necessidades e desafios específicos de cada região.

Ao combinar as informações dos gráficos anteriores, é possível compreender a produção de cada estado em relação aos tipos de crimes. A **Figura 2**, a seguir, apresenta a composição dos resultados.

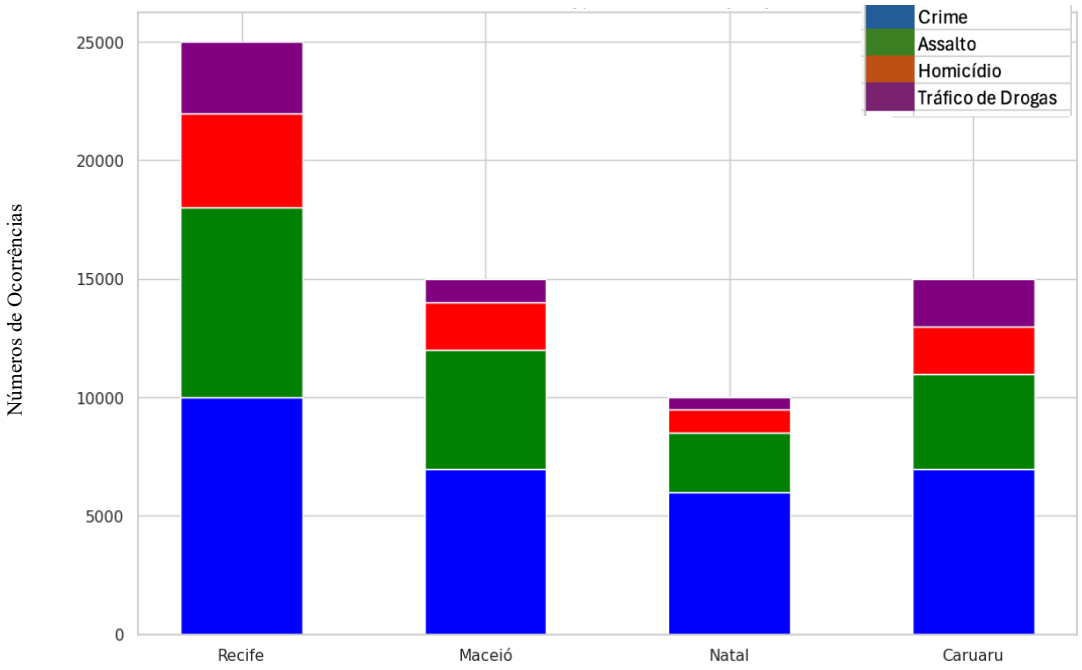

Figura 2**.** Distribuição de termos relacionados à violência

A Figura 2 mostra a distribuição de vários tipos de violência—crime, agressão, homicídio e tráfico de drogas—nas cidades de Recife, Maceió, Natal e Caruaru. No entanto, é importante destacar que essa análise está baseada apenas nos dados coletados, sendo necessário considerar outras fontes e análises complementares para uma compreensão mais abrangente do cenário.

Maceió registra 7.000 ocorrências de crimes, influenciadas pela urbanização e fatores socioeconômicos. A cidade também apresenta um alto número de agressões, com 5.000 incidentes registrados. Os homicídios, totalizando 2.000 ocorrências e o tráfico de drogas é menos prevalente, com 1.000 ocorrências, em comparação com Recife.

Em Natal, há 6.000 ocorrências de crimes gerais. As agressões, com 2.500 ocorrências, são significativas, mas menores do que em Recife e Maceió. A cidade registra 1.000 homicídios, indicando um controle relativamente melhor sobre os crimes violentos. O tráfico de drogas é o mais baixo entre as cidades, com 500 ocorrências.

Caruaru registra 7.000 ocorrências de crimes gerais, semelhantes a Recife, refletindo problemas socioeconômicos. As agressões são uma grande preocupação, com 4.000 incidentes registrados. Os homicídios, totalizando 2.000 ocorrências, destacam sérias preocupações com crimes violentos. A cidade também registra 2.000 incidentes de tráfico de drogas.

Desta forma, o estudo identificou 65.000 artigos de notícias relacionados à violência em quatro cidades brasileiras: Recife, Maceió, Natal e Caruaru. Para dimensionar a relevância desses dados, é pertinente compará-los com as taxas nacionais de homicídios.

Segundo o Atlas da Violência 2023, o Brasil registrou uma taxa de 21,1 homicídios por 100 mil habitantes em 2022, totalizando 45.474 assassinatos. Já o Atlas da Violência 2024 indica uma média de 21,7 homicídios por 100 mil habitantes em 2022, podendo chegar a 24,5 quando considerados os homicídios ocultos.

Ao analisar os dados específicos das cidades:

- Recife registrou 25.000 artigos, representando quase 40% do total, o que evidencia uma cobertura midiática intensa e possivelmente uma elevada incidência de violência.
- Maceió e Caruaru, com 15.000 artigos cada, apresentam números significativos, especialmente considerando suas populações menores em comparação a Recife.
- Natal, com 10.000 artigos, também demonstra uma presença considerável de notícias relacionadas à violência.

A frequência de termos como "crime" (30.000 ocorrências), "assalto" (15.000), "homicídio" (10.000) e "tráfico de drogas" (10.000) reforça a ideia de que a violência está disseminada e assume múltiplas formas nessas localidades.

Esses dados, quando comparados às taxas nacionais de homicídios, sugerem que as cidades analisadas enfrentam desafios significativos em termos de segurança pública. A elevada quantidade de artigos e a diversidade dos crimes reportados indicam a necessidade de estratégias específicas e direcionadas para cada região, considerando seus contextos socioeconômicos e particularidades locais.

3.1 Implicações do estudo

O estudo destaca a necessidade de estratégias de segurança pública. Programas de envolvimento comunitário são fundamentais para fortalecer a confiança e promover a resolução de conflitos. Medidas preventivas, como campanhas educativas e iniciativas de saúde mental, podem abordar questões subjacentes que contribuem para a violência. Pesquisas contínuas e monitoramento são essenciais para acompanhar a eficácia das estratégias e ajustá-las com base em novas tendências.

Além disso, este estudo traz relevância para a sociedade, pois fornece um método que permitiu realizar uma análise preliminar sobre os padrões de violência. A utilização da computação no processamento e análise desses dados possibilita a identificação de padrões, tendências e áreas críticas de forma rápida e precisa. O uso de tecnologias avançadas, como inteligência artificial e big data, pode otimizar a alocação de recursos e fortalecer medidas preventivas, contribuindo para uma gestão mais eficiente da segurança pública.

Essa abordagem busca aprimorar a segurança pública e melhorar a qualidade de vida dos moradores dessas cidades, demonstrando como a combinação de análise de dados e

computação pode ser uma ferramenta interessante na formulação de políticas públicas mais eficazes

5. Conclusão

Este estudo forneceu contribuições relevantes ao mapear a distribuição e a intensidade de diferentes tipos de violência em Recife, Maceió, Natal e Caruaru, a partir de um corpus de notícias coletadas ao longo de dois anos. A delimitação temporal confere maior consistência às análises, permitindo identificar padrões relevantes de criminalidade dentro de um intervalo definido, o que fortalece a sustentação das recomendações apresentadas.

Os dados indicam que há, de fato, variações significativas entre as cidades quanto à incidência de crimes como homicídios, tráfico de drogas e agressões, o que evidencia a necessidade de estratégias de segurança pública adaptadas aos contextos locais. Ainda assim, é necessário adotar uma postura crítica quanto à abrangência e à natureza das fontes utilizadas: como se trata de um corpus derivado de notícias jornalísticas, os dados podem refletir vieses de cobertura e não necessariamente representar a totalidade dos crimes ocorridos. Para mitigar essas limitações, o trabalho pretende incorporar fontes complementares, como dados oficiais de órgãos de segurança pública, registros policiais e estatísticas governamentais. Além disso, a adoção de técnicas de validação cruzada entre diferentes bases de dados pode ajudar a corrigir distorções e aumentar a confiabilidade das análises. Futuramente, planeja-se integrar análises geoespaciais e realizar estudos longitudinais mais amplos para captar tendências temporais e espaciais da violência. O envolvimento comunitário, por meio de pesquisas participativas, também será explorado como uma forma de enriquecer os dados qualitativos e captar percepções locais que nem sempre são refletidas pela mídia.

A análise evidencia o potencial do uso de tecnologias avançadas, como inteligência artificial e big data, na construção e exploração de grandes volumes de dados textuais. No entanto, a eficácia dessas ferramentas depende do rigor metodológico e da integração com outras bases de dados oficiais. Direções futuras incluem a realização de estudos longitudinais mais extensos, comparações regionais ampliadas, análises geoespaciais e iniciativas de pesquisa participativa com envolvimento comunitário.

A partir dessas abordagens, será possível avançar na formulação de políticas públicas baseadas em evidências, mais eficazes no enfrentamento da violência urbana e na promoção da qualidade de vida da população.

Agradecimentos

CNPQ (Conselho Nacional de Desenvolvimento Científico e Tecnológico). Processo: 409237/2022-1. Ação: Chamada CNPq/SEMPI/MCTI/FNDCT Nº 54/2022

Referências

Carnaz, G.; Antunes, M.; Nogueira, V.B. An Annotated Corpus of Crime-Related Portuguese Documents for NLP and Machine Learning Processing. Data 2021, 6(7), 71.

de Carvalho, V.D.H.; Costa, A. Towards Corpora Creation from Social Web in Brazilian Portuguese to Support Public Security Analyses and Decisions. *Libr. HI TECH* 2022, doi:10.1108/LHT-08-2022-0401.

de Carvalho, V.D.H.; Nepomuceno, T.C.C.; Poleto, T.; Turet, J.G.; Costa, A.P.C.S. Mining Public Opinions on COVID-19 Vaccination: A Temporal Analysis to Support Combating Misinformation. *Trop. Med. Infect. Dis.* 2022, *7*, 256, doi:10.3390/tropicalmed7100256.

de Carvalho, V. D. H., Costa, A. P. C. S.(2023). Exploring Text Mining and Analytics for Applications in Public Security: an in-depth dive into a Systematic Literature Review.Socioeconomic Analyt-ics,1(1), 5-55.

FÓRUM BRASILEIRO DE SEGURANÇA PÚBLICA. Atlas da Violência 2023. São Paulo: FBSP; Ipea, 2023. Disponível em: https://www.ipea.gov.br/atlasviolencia/. Acesso em: 15 maio 2025.

FÓRUM BRASILEIRO DE SEGURANÇA PÚBLICA. Atlas da Violência 2024. São Paulo: FBSP; Ipea, 2024. Disponível em: https://www.ipea.gov.br/atlasviolencia/. Acesso em: 15 maio 2025.

FÓRUM BRASILEIRO DE SEGURANÇA PÚBLICA. Atlas da Violência 2025. São Paulo: FBSP; Ipea, 2025. Disponível em: https://www.ipea.gov.br/atlasviolencia/. Acesso em: 15 maio 2025.

Gomes, R.; Ferreira, S.; Almeida, P. Using Machine Learning to Detect Cyberbullying in Social Media: A Study in Portuguese. Cyberpsychol. Behav. Soc. Netw. 2021, 24, 78, doi:10.1089/cyber.2020.0456

Silva, J.R.; Santos, M.P.; Oliveira, L.F. Sentiment Analysis of Social Media Data for Crime Prediction: A Case Study in Brazil. J. Comput. Sci. 2021, 15, 45, doi:10.3390/jcs15020045

Suhaimin, M.S.M.; Hashim, H.; Zainol, Z.; Chien, S.F. Social Media Sentiment Analysis and Opinion Mining in Public Security: Taxonomy, Trend Analysis, Issues and Future Directions. Libr. HI TECH 2022, doi:10.1108/LHT-08-2022-0401.

Turet, J.; Costa, A.P.C.S. Hybrid Methodology for Analysis of Structured and Unstructured Data to Support Decision-Making in Public Security. Data 2022, 6, 91, doi:10.3390/data06010091.