Extração e Avaliação de uma Base de Dados sobre Criminalidade em Português a partir do Twitter

Gabriel V. da Fonseca Miranda[1], **Vinícius Gabriel de J. Almeida**[1], **Thais R. M. Braga Silva**[1], **Fabrício A. Silva**[1]

[1]Laboratório de Inteligência em Sistemas Pervasivos e Distribuídos (NESPeD-Lab)
Instituto de Ciências Exatas e Tecnológicas (IEF).
Universidade Federal de Viçosa (UFV) - Florestal – MG – Brasil

{gabriel.v.miranda, vinicius.jesus, thais.braga, fabricio.asilva}@ufv.br

Abstract. *In the last years studies on security solutions to smart homes, transportation systems and even cities have been developed. In this scenario, criminal data have became increasly important. Although occurences from police databases are frequently used, many times the most ordinary crimes end up not being registered by them. The goal of this work is to present a method to extract criminal data from São Paulo city portuguese Twitter posts. The majority of related work found perform automatic extractions for english texts. When portuguese is considered, the accurancy is frequently not presented and the final dataset is small. In this work, the final dataset has 1,333 labeled tweets, which were compared to a police database, highlighting information similarities but also possibilities for complementation.*

Resumo. *Nos últimos anos, trabalhos que descrevem soluções de segurança para casas, sistemas de transporte e até mesmo cidades inteligentes têm sido desenvolvidos. Neste cenário, dados sobre criminalidade têm se tornado cada vez mais importantes. Embora ocorrências em bases policiais sejam frequentemente utilizadas, muitas vezes os crimes mais corriqueiros acabam não sendo registrados dessa forma. O objetivo deste trabalho é apresentar um método para a extração de dados de criminalidade em português a partir de postagens no Twitter feitas na cidade de São Paulo. A maioria dos trabalhos relacionados encontrados faz esse tipo de extração automatizada para textos em inglês. Quando o português é considerado, frequentemente a acurácia não é apresentada e a base final é pequena. Neste trabalho, a base de dados final possui 1.333 tweets rotulados, que foram comparados a uma base policial, mostrando similaridades e possibilidades de complementação de informações.*

1. Introdução

Nos últimos anos a comunidade de pesquisa na área de computação ubíqua e pervasiva tem desenvolvido trabalhos que visam oferecer soluções de segurança para casas, sistemas de transporte e até mesmo cidades inteligentes [Adesola et al. 2019]. De modo geral, as aplicações ligadas a estes trabalhos visam obter informações sobre situações ou lugares potencialmente perigosos e utilizá-las para aumentar o nível de segurança de seus usuários [Laufs et al. 2020, Sarhan 2020, Neto et al. 2018].

Apesar de diferentes entre si, várias dessas aplicações enfrentam um desafio em comum, qual seja, a obtenção de bases de dados sobre criminalidade que sejam não só

volumosas, mas também completas e coerentes com os propósitos almejados. A obtenção de dados em quantidade e qualidade, de modo geral, pode ser considerada uma tarefa desafiadora [Cai and Zhu 2015]. É muito frequente que bases de dados contenham ruídos, dados duplicados ou ainda incompletos. Além disso, muitas vezes os dados não estão disponíveis de forma pública, nem mesmo para pesquisas científicas.

Os dados sobre criminalidade utilizados para realização destes tipos de trabalho são, frequentemente, obtidos a partir de boletins de ocorrência (BOs). Isso porque as secretarias de segurança de algumas cidades, tais como São Paulo/SP, disponibilizaram nos últimos anos esses registros para estudos científicos de forma anonimizada e estruturada[1]. Porém, ainda neste caso, apesar do acesso e da grande quantidade de registros, muitas vezes os relatos de crimes que seriam do interesse das aplicações ubíquas e pervasivas, tais como roubo de celulares, furtos de objetos de pequeno valor e assédio, não estão disponíveis, uma vez que a população acaba por não registrá-los junto a polícia[2].

Uma hipótese possível sobre como obter mais dados de criminalidade, e, em especial, estes que acabam não sendo registrados, seria buscando-os em relatos postados por usuários de redes sociais. Dado que no Brasil um grande número de pessoas de modo geral possuem cadastros em sistemas desse tipo e fazem uso constante dos mesmos, realizando postagens inclusive com relatos pessoais e situações do dia a dia [3], é possível supor que registros vivenciados ou testemunhados de situações criminais possam ser encontrados. Vários desafios, no entanto, estão ligados a possibilidade de extração desse tipo de dado. Um deles é a necessidade de que a localização do registro criminal também esteja disponível, uma vez que não basta saber qual crime ocorreu mas também onde. Além disto, a extração de dados a partir de textos na língua portuguesa pode ser ainda mais complicada, especialmente devido às ambiguidades no uso de palavras e termos. Em particular, o vocabulário utilizado para a descrição de crimes é muito recorrente também em situações diversas, o que dificulta o processo tornando-o menos preciso.

O objetivo, portanto, deste trabalho, é apresentar um método para a extração de dados de criminalidade em português a partir da rede social Twitter. Como estudo de caso será considerada a cidade de São Paulo, escolhida por ser a maior metrópole brasileira e, portanto com maior população, grande quantidade de usuários da rede social utilizada e alto volume de situações de insegurança. Porém, os passos realizados podem também ser aplicados para outras cidades. O Twitter foi escolhido uma vez que, para trabalhos acadêmicos, é possível ter acesso a sua base de dados completa via uma API de consulta. Além disso, nesta rede os usuários costumam realizar postagens com frequência, muitas vezes contendo suas localizações geoespaciais e relatando acontecimentos do cotidiano. A maioria dos trabalhos relacionados encontrados na literatura faz extração automatizada de dados sobre criminalidade para textos em inglês. Quando o português é considerado, frequentemente a acurácia não é apresentada e a base final é pequena. Além do método apresentado em si, a base de dados construída a partir dele para a cidade de São Paulo está também disponibilizada como uma contribuição deste trabalho[4].

---

[1] http://www.ssp.sp.gov.br/transparenciassp/

[2] https://agenciabrasil.ebc.com.br/geral/noticia/2022-12/pesquisa-do-ibge-mostra-subnotificacao-de-roubos-e-furtos-no-brasil

[3] https://resultadosdigitais.com.br/marketing/estatisticas-redes-sociais/

[4] https://github.com/NESPEDUFV/repositorio_dados_sbcup

Neste trabalho, também será realizada uma análise exploratória de dados, utilizando os conjuntos de dados criminais do Twitter e os boletins de ocorrência de São paulo. Esses dados serão utilizados para identificar as características mais relevantes do crime, e para fazer previsões. Para isso, serão utilizados dois classificadores, o Naive Bayes e o Árvore de Decisão. Esses classificadores serão utilizados para realizar previsões de tipos de crimes na cidade de São Paulo.

O restante deste artigo está organizado da seguinte forma: a Seção 2 apresenta trabalhos que abordam a extração de dados de criminalidade a partir do Twitter; a Seção 3 apresenta o passo a passo do método aplicado para construção da base de dados de criminalidade a partir do Twitter para a cidade de São Paulo; Uma análise comparativa dos dados extraídos via Twitter com aqueles disponíveis em uma base de dados de boletins de ocorrência policiais é apresentada na Seção 4; a Seção 5 apresenta uma análise exploratória dos dados provenientes do Twitter e da Polícia, com o objetivo de realizar previsões sobre os tipos de crimes que ocorrem nos diversos bairros da cidade de São Paulo; Por fim, algumas conclusões e apontamentos de possíveis trabalhos futuros podem ser encontrados na Seção 6.

2. Trabalhos Relacionados

O trabalho de [Clarindo et al. 2016] apresenta o sistema DETECT, cujo objetivo é utilizar análise de sentimentos em tweets com foco em encontrar postagens contendo relatos criminais, principalmente relacionados a crimes patrimoniais. Os autores utilizaram dados obtidos de fontes externas ao Twitter. No total, o sistema detectou 569 tweets geolocalizados relevantes no contexto de violência patrimonial. Um mapa de calor foi criado para demonstrar as áreas da cidade do Rio de Janeiro com as maiores ocorrências de crimes no período de setembro de 2015 a janeiro de 2016. No artigo de [Prathap and Ramesha 2018], também foi utilizada a análise de sentimentos, porém em tweets coletados por meio da API versão 1 disponibilizada pela plataforma Twitter. Os resultados obtidos foram apresentados em gráficos que permitem visualizar a quantidade total de palavras relacionadas a crimes de roubo.

Em [Secron et al. 2016] é proposta a ferramenta SigaCiente, que utiliza dados do Twitter, obtidos por meio da API versão 1, para adicionar informações de criminalidade às rotas de trânsito na cidade do Rio de Janeiro. Com uma coleta de dados com duração de uma semana, dos cerca de 7.000 tweets obtidos apenas 500 foram selecionados para a análise. De acordo com os autores, a ferramenta obteve uma margem de acerto de 95% na identificação de rotas seguras e inseguras.

O trabalho de [dos Santos 2015] utiliza a API 1.1 do Twitter para capturar tweets contendo informações criminais no estado de São Paulo. Foram selecionadas as palavras-chave como agressão, assalto, roubo, vítima, morte, entre outras, e coletados mais de 50 mil tweets durante 9 dias no mês de setembro. Para classificar os dados obtidos, foram utilizados e comparados diferentes algoritmos de aprendizado de máquina.

Considerando postagens em inglês, [Gerber 2014] criou um modelo de previsão de crimes para a cidade de Chicago/EUA, a partir de dados coletados de uma plataforma externa ao Twitter, entre janeiro e março de 2013. Foram coletados 1.528.184 de tweets com informações de geolocalização a partir dos quais foi criado um modelo de previsão baseado em tópicos. Já [dos Reis and Nakamura 2017] teve como objetivo criar uma base

de dados com relatos criminais relacionados a crimes na cidade de Nova Iorque/EUA, utilizando a API do Twitter para capturar tweets recentemente publicados. Foram obtidos 316.977 relatos criminais que ocorreram entre os meses de maio e dezembro de 2016. Os tweets foram rotulados com nove categorias e utilizados para construção de um modelo de classificação. O trabalho [Mahajan and Mansotra 2021] usa dados do Twitter geolocalizados de cinco regiões da Índia para prever crimes. Para isso, uma técnica de análise semântica de sentimentos foi utilizada usando uma arquitetura bidirecional de memória longa-curta (BiLSTM). Os tweets processados alimentam uma rede neural para determinar a intensidade do crime em uma determinada área. A eficácia dessa abordagem foi de 84.74% para as classes de sentimento. Os resultados também mostraram uma correlação entre os padrões de crime previstos por tweets e crimes reais. No artigo de [Vivek and Prathap 2023] também foi utilizada a API do Twitter para análise e previsão de crimes na Índia. Cinco estados indianos foram analisados, e tweets relacionados a crimes foram extraídos do Twitter durante 2 meses. Após a limpeza dos dados, 15.601 tweets foram usados. O trabalho comparou três modelos de previsão de séries temporais: ARIMA, SARIMA e LSTM, sendo o primeiro escolhido como o mais adequado por apresentar o menor RMSE (*Erro Médio de Raiz Quadrada*).

O artigo [Almanie et al. 2015] apresenta modelos de previsão de crimes para duas cidades, com base em características como mês, dia, hora e local dos crimes. O objetivo é conscientizar as pessoas sobre áreas perigosas. Para isso foram utilizados dois classificadores: Naive Bayes e Árvore de Decisão, para realizar as previsões de crimes em Denver e Los Angeles. O classificador Naive Bayes obteve uma acurácia de 51% para Denver e 54% para Los Angeles, enquanto o classificador Árvore de Decisão apresentou uma acurácia de 42% para Denver e 43% para Los Angeles.

Em [Khan et al. 2022], é apresentado um modelo de previsão de crimes que utiliza três algoritmos de classificação: Naive Bayes, Floresta Aleatória e Árvore de Decisão. Esses classificadores são aplicados para analisar tendências de crimes e realizar previsões sobre diferentes categorias de crimes na cidade de São Francisco. As acurácias obtidas para cada modelo foram de 65,82% para Naive Bayes, 63,43% para Floresta Aleatória e 98,5% para Árvore de Decisão. O objetivo desse trabalho é auxiliar os órgãos de segurança a antecipar ocorrências criminais em momentos específicos. Isso pode contribuir para a implementação de medidas preventivas mais eficazes.

Os trabalhos acima apresentam soluções que visam extrair informações criminais a partir de dados do Twitter. Neste trabalho, ao contrário dos demais, com uso da API versão 2 do Twitter foi feita uma busca por dados geolocalizados sobre crimes em mensagens em português, considerando a cidade de São Paulo. Os dados foram manualmente analisados e classificados, gerando uma base confiável de 1.333 tweets com informações criminais. Ademais, foi feita uma análise comparativa dos dados gerados com uma base de boletins de ocorrência disponibilizada pela polícia. Por fim, também foi realizado um previsão de tipos de crimes usando ambas as bases de dados. Para essa previsão foram usados dois classificadores o Naive Bayes e o Árvore de Decisão.

3. Extração de Dados de Criminalidade no Twitter

O objetivo deste trabalho é utilizar o Twitter como fonte alternativa de dados para identificação de registros de ocorrências criminais. Para isso, foi feita uma busca por

informações criminais a partir de tweets obtidos por meio de acesso a esta rede social via API própria, culminando na construção de uma base de dados que possa ser utilizada para a identificação mais completa de áreas criminais existentes na cidade.

3.1. Uso da API para Coleta de Dados da Base do Twitter

Na busca por informações de crimes em tweets, foi possível usar, neste trabalho, a API versão 2 do Twitter [5], disponível a partir de um cadastro para projetos acadêmicos. Essa API permite acessos mais avançados a uma maior quantidade de dados da plataforma, possibilitando resgate de até 10 milhões de tweets por mês. Como era importante que os tweets obtidos possuíssem sua geolocalização (latitude e longitude), uma vez que essa característica foi utilizada para estabelecer a localização para o crime, e observando que poucos usuários do Twitter na cidade de São Paulo compartilham essa informação, foi necessário utilizar um intervalo temporal amplo, resgatando da rede social dados entre 2010 e 2022, de modo a se obter um maior volume retornado. Além da *geolocalização*, os atributos *idtweet* e *texto* foram selecionados para serem retornados pela API. A estrutura utilizada para o armazenamento de cada tweet com informação criminal é composta, portanto, por estes campos, além do campo *rótulo*, que indica o tipo de crime associado. A forma como este campo é preenchida será explicada na seção 3.3.

Um impasse na versão 2 desta API é que a mesma não possui uma funcionalidade para se obter dados a partir de uma dada latitude e longitude, passando-se, a partir delas, um raio para abranger uma certa área. Desta forma, foi necessário inicialmente obter todos os tweets com geolocalização no território brasileiro. Em seguida, para filtrá-los, foram utilizadas duas estratégias que envolvem o uso de criação de polígonos que representem regiões. Primeiro, foi feita uma análise mais geral, para checar se o crime está dentro da cidade ou não. Desta forma, foi utilizada a biblioteca *"OSMnx: Python for street networks"* [6], disponível como um pacote da linguagem python para a criação de grafos de cidades. Entretanto, para essa etapa de checagem, não foi necessário criar efetivamente o grafo. O uso da funcionalidade *geocode_to_gdf* [7], disponível na biblioteca OSMnx, foi suficiente para construir o polígono da cidade. Por se tratar de uma função que retorna um polígono no formato da classe *Polygon*, disponível na biblioteca *Shapely* [8] do Pyhton, fez-se a simples checagem da geolocalização dos crimes, utilizando o método *contains*, que retorna se uma coordenada está localizada no polígono ou não. Com isso, foi necessário apenas passar pelos tweets e checar as coordenadas com este método.

A segunda etapa implementada foi calcular a distribuição dos crimes pelos bairros da cidade. O método utilizado para identificar se certo crime estava dentro de um bairro ou não foi idêntico à etapa anterior, criando polígonos com o auxílio da biblioteca *Shapely*. A única diferença presente no processo, foi a origem dos dados utilizados para criar os polígonos, pois, por se tratar de um nível de abstração maior, sendo necessário obter a divisão entre bairros da cidade, a base de dados [9] da DataGeo [10] foi utilizada. Na

[5] https://developer.twitter.com/en/docs/twitter-api

[6] https://osmnx.readthedocs.io/en/stable/

[7] https://osmnx.readthedocs.io/en/stable/osmnx.html#osmnx.geocoder.geocode_to_gdf

[8] https://shapely.readthedocs.io/en/stable/manual.html#polygons

[9] https://encurtador.com.br/wxKRT

[10] https://datageo.ambiente.sp.gov.br/

plataforma, foi possível extrair os dados no formato *geojson*, compatível com a biblioteca *Shapely*. Por fim, bastou fazer a mesma iteração com os tweets e identificar em qual bairro cada crime estava localizado. A partir desta implementação, o atributo *bairro* pode ser adicionado a base de dados de tweets criminais.

3.2. Definição de Palavras Chaves sobre Criminalidade

Para selecionar apenas tweets contendo as palavras-chave consideradas como relacionadas ao vocabulário de relatos criminais, também foi utilizado um filtro. Os trabalhos relacionados descritos na Seção 2 e que também consideraram tweets em português foram utilizados como referência, apresentando uma série de possíveis termos provavelmente relacionados a algum relato criminal, tais como: roubo, homicídio, assédio, assalto, dentre outros. Outras palavras também foram selecionadas empiricamente para se fazer a recuperação de dados pela API do Twitter, sendo ao todo escolhidas 23 palavras-chave. Vale destacar aqui que as especificidades do português quanto à determinação de gênero e tempo verbal foram dificultadores neste processo. Muitos dos termos relacionados possuem versões no masculino e feminino (e.g., roubado e roubada), além de serem mencionadas em conjugações distintas (e.g., roubou, roubando, roubado). Assim, foi necessário considerar, sempre que possível, apenas o radical (e.g., roub), o que aumenta o risco de serem selecionados tweets não relacionados.

3.3. Limpeza, Rotulagem e Armazenamento de Tweets Resultantes

Em um primeiro momento, foram retornados 12.024 tweets geolocalizados dentro da cidade de São Paulo, no intervalo temporal determinado e contendo as palavras-chave escolhidas. Os tweets foram armazenados em uma base de dados MongoDB. Entretanto, ao inspecionarmos esse resultado, foi possível observar que muitos dos tweets obtidos não tinham relação com informações criminais. Isso se deve ao fato de que, no português, é muito comum o uso do vocabulário criminal em diversas outras circunstâncias, como políticas, esportivas ou até mesmo afetivas. Desta forma, ficou clara a necessidade de se realizar uma limpeza. Inicialmente, foi feita uma análise para encontrar padrões de palavras inseridas nos textos de tweets (e.g., política, juiz, coração), a partir das quais fosse possível identificar e remover de forma automática postagens que não estavam informando um crime. Entretanto, essa abordagem pode eliminar tweets válidos, além de não resolver o problema de forma definitiva.

Dessa forma, para melhorar a qualidade da base e ainda possibilitar a rotulagem dos tweets com o tipo de crime por ele reportado, foi criada uma aplicação Web. O processo de desenvolvimento foi dividido em duas partes principais: uma API como *back-end* , e uma página Web como *front-end* . A API foi construída utilizando a biblioteca NodeJS em conjunto com o banco de dados MongoDB. O objetivo principal de combinar essas ferramentas foi para facilitar a disponibilização de *endpoints*, a fim de permitir de maneira simples a inserção, alteração e extração dos dados. Com a estrutura da API montada, bastou criar um cliente Web para facilitar a comunicação com os *endpoints* e tornar a tarefa de rotulagem menos exaustiva. Para o seu desenvolvimento foi utilizada a biblioteca de *front-end* ReactJS, a linguagem de estilização Sass, e a biblioteca de componentes Material UI. O resultado final deste projeto está em um repositório público do GitHub [11].

[11] `https://github.com/NESPEDUFV/rotulagem_tweets_nesped_ufv`

As Figuras 1a e 1b mostram a interface da página web criada. Percebe-se que o uso de botões com as categorias de interesse mostrou-se versátil tanto para facilitar a iteração do usuário com a página, quanto para restringir as opções de rótulos. Foram criados botões únicos para crimes como: furto, roubo, assédio, entre outros, pois essas são categorias bem distintas e com frequência maior de registros e ocorrências. Desta forma, categorias menos comuns foram abrangidas na opção "Outros". A partir desta rotulagem, o atributo *categoria do rótulo* pode ser adicionado a base de dados de tweets criminais.

(a) Tweet com registro criminal **(b) Tweet sem registro criminal**

Figura 1. Página Web para rotulagem dos tweets criminais.

Após a utilização do sistema web explicado acima, e analisando-se portanto manualmente e individualmente os 12.024 tweets inicialmente obtidos, foram selecionados e rotulados ao final 1.333 tweets contendo um relato criminal. Este resultado mostra que apenas 11% dos tweets resgatados eram de fato relatos criminais, o que ressalta como o vocabulário criminal em português é de fato muito utilizado em diversas outras situações.

4. Análise: Dados do Twitter x Dados Policiais

Com o objetivo de analisar os dados de criminalidade extraídos do Twitter, algumas comparações foram feitas considerando a base de dados construída com os mesmos e uma outra composta por boletins de ocorrência registrados na cidade de São Paulo. Essa comparação permite a observação de semelhanças e diferenças entre elas, de maneira a verificar como podem se complementar em aplicações de segurança ubíquas e pervasivas.

A base de dados de boletins de ocorrência de São Paulo foi obtida por meio do Portal de Transparência da cidade [12]. Os dados criminais estão, a princípio, listados em categorias de crimes mais abrangentes. Para este trabalho, foram utilizadas as categorias: *furto e roubo de celular* e *furto e roubo de veículos*. Dentro de cada categoria, é possível separar os crimes por ano e mês. Desta forma, basta escolher o recorte temporal e baixar os arquivos dos meses separadamente. Para este trabalho, foram escolhidos todos os meses do ano de 2019. Esses dados estão no formato de planilhas (.xls) e possuem 54 colunas que descrevem cada ocorrência registrada. A base passou inicialmente por um processo de limpeza para que apenas dados válidos para o trabalho pudessem ser considerados. Desta forma, apenas dados com geolocalização foram aproveitados, levando a uma base de dados geolocalizados com aproximadamente 310 mil crimes e 12 colunas. A redução das colunas foi feita para deixar apenas as informações principais dos crimes,

[12] `http://www.ssp.sp.gov.br/transparenciassp/`

como data e hora, latitude e longitude, bairro, e rubrica, responsável por descrever o tipo do crime cometido. Além disso, apesar de realizar o filtro dos crimes geolocalizados, muitas rubricas presentes na base não tinham muita relação com o quesito mobilidade e segurança urbana. Com isso, foi realizado outro filtro, deixando a base ao final com aproximadamente 212 mil registros criminais geolocalizados.

Uma vez que as duas bases de dados possuem informações de criminalidade geo-referenciadas, foi realizada a plotagem destas utilizando-se a técnica de mapa de calor. A figura 2a apresenta o mapa para a base criminal com dados da base da polícia completo, enquanto na figura 2b estão os crimes que advém dos tweets. Uma terceira base de dados foi criada a partir da filtragem sobre os dados de boletins de ocorrência, de maneira a manter apenas aqueles cujos tipos de crimes associados fossem semelhantes aos considerados para a base do Twitter. Dos 222 tipos de crimes iniciais da base policial foram mantidos apenas 89. A figura 2c apresenta o mapa de calor da base policial filtrada. Observando as figuras, é notável que a quantidade de crimes plotados usando o mapa de calor é superior ao da base de dados do Twitter. Porém, no geral os 3 mapas conseguem cobrir grande parte do município de São Paulo. Um ponto em especial é a região central do município, que aparece bem mais densa para todos, fortalecendo assim a existência de um grande quantidade de ocorrências de crimes nesta região.

(a) Policial Completa **(b) Twitter** **(c) Policial Filtrada**

Figura 2. Mapa de Calor - Registros de Crimes/São Paulo

(a) Policial Completa **(b) Twitter** **(c) Policial Filtrada**

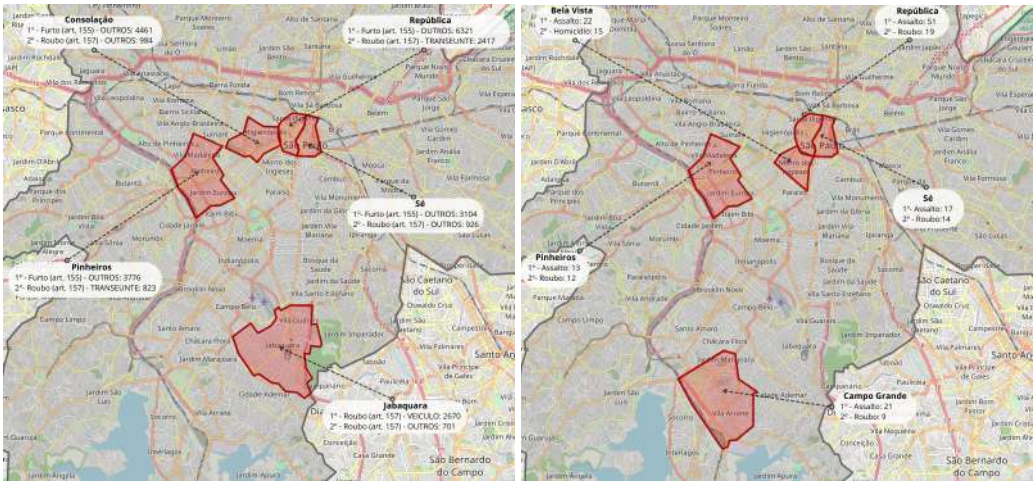
Figura 3. Mapa de Calor - Registros de Crimes/Zoom

As figuras 3a, 3c e 3b apresentam uma ampliação do mapa de calor para as 3 bases. É possível observar que com os dados da base policial completa, a cidade fica completamente tomada por informações de criminalidade, o que torna difícil para as aplicações estabelecerem prioridades ou encontrar alternativas sobre uso seguro de espaços. A base de dados de tweets pode ser utilizada como uma referência neste caso.

Na base de dados da polícia, os 3 tipos de crimes mais recorrentes são os seguintes: *Furto (art. 155) - OUTROS* ocupa o primeiro lugar com um total de 60.025 crimes, seguido por *Roubo (art. 157) - VEICULO* em segundo lugar com 46.054 crimes e *Roubo (art. 157) - TRAUSEUNTE* em terceiro lugar com 42.654 crimes relatados a polícia. Já na

base de dados do Twitter a liderança é do tipo de crime *Assalto* com 752 crimes, seguido por *Roubo* em segundo lugar com 339 crimes, e *Homicídio* em terceiro lugar com um total de 76 crimes rotulados.

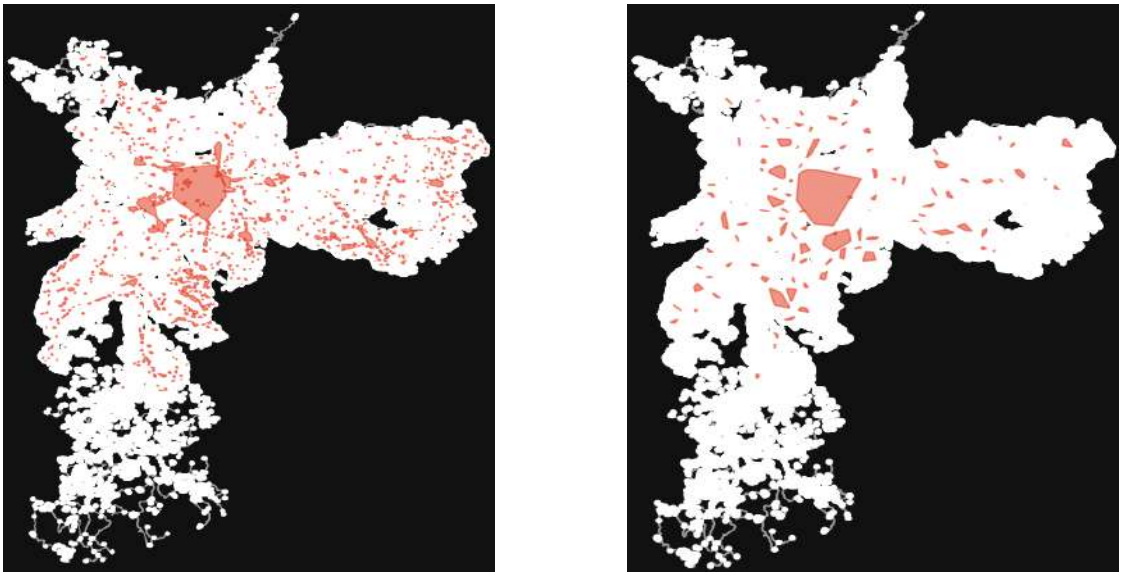
As figuras 4b e 4a apresentam no mapa da cidade de São Paulo os 5 bairros mais perigosos das duas bases e seus dois tipos de crimes mais comuns. É possível notar como as bases apontam para informações geolocalizadas de crimes parecidas, porém complementares. No bairro República, por exemplo, o principal tipo de crime relatado é diferente para as duas bases (furto para a policial e assalto para o Twitter). A base do Twitter indica que o bairro Bela Vista, próximo aos bairros República e Sé, também apresentam índices de criminalidade relevantes.



(a) Policial Filtrada **(b) Twitter**

Figura 4. Bairros mais perigosos por base de dados

Por fim, como um dos principais objetivos a partir do levantamento de bases de dados criminais é a identificação de áreas potencialmente perigosas na cidade, as duas bases de dados foram utilizadas junto ao algoritmo de clusterização DBScan. Neste caso, foi considerada a base policial completa. Além disso, os seguintes valores foram utilizados para os parâmetros número de vizinhos (υ) e distância máxima ε entre pontos em cada caso: na base de dados policiais υ = 36 e ε = 100; já na base de dados do Twitter υ = 3 e ε = 600. Os valores foram determinados após a realização de análises exploratórias. As figuras 5a e 5b apresentam os resultados da criação dos clusters para a base de dados policial e a de tweets, respectivamente. É possível observar que o padrão de criação das áreas criminais é muito semelhante, com as áreas mais perigosas podendo ser ressaltadas pela base de dados do Twitter. Foram gerados 987 e 124 clusters para a base da polícia e do Twitter, respectivamente. No primeiro caso, os maiores clusters estavam nos bairros República, Pinheiros e Santana. Já no segundo, Bela Vista, Saúde e Campo Grande. Os bairros República e Pinheiros são vizinhos no centro de São Paulo. Por fim, é interessante ressaltar como os padrões de localização de crimes se mantém em todas as formas de visualização mostradas neste trabalho.



(a) Base Policial (b) Base Twitter

Figura 5. Clusters obtidos com uso do DBScan.

5. Previsão de Crimes

Nesta parte, serão analisados os principais atributos das bases de dados para encontrar associações entre as características do crime. O foco dessa abordagem será verificar se ocorrem mais crimes durante feriados e finais de semana ou durante a semana, quais dias e meses do ano são os mais perigosos e quais dias da semana e horários durante o dia apresentam maior risco de crime. Essa abordagem visa extrair padrões de atributos do crime. Posteriormente, serão utilizados algoritmos de classificação para prever tipos de crimes em um determinado local, a fim de evitar situações de risco. Para essa etapa, serão considerados 728 tweets e 214.373 boletins de ocorrência da base da Polícia filtrada, pois somente estes possuem os atributos de data e hora. A análise feita aqui tem como base o seguinte repositório [13].

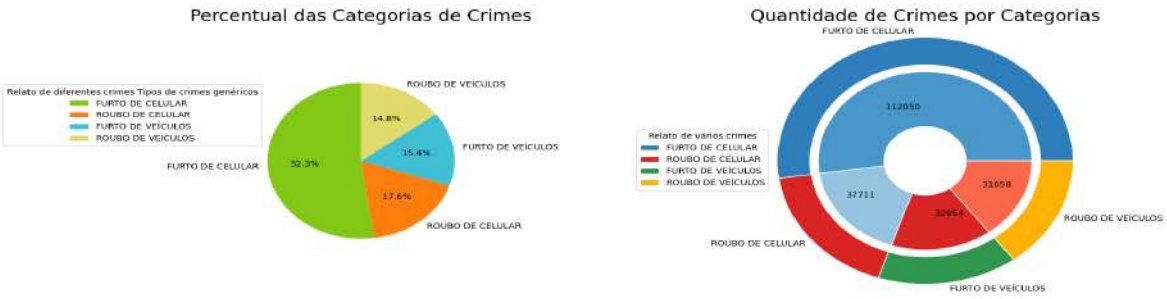
5.1. Análise Exploratória de Dados

Como passo inicial para a análise exploratória dos dados, foram realizadas análises dos atributos dos conjuntos de dados. Então, para a base de dados da polícia filtrada e do Twitter, foram gerados diversos gráficos com o objetivo de aprofundar a compreensão dos dados. Os gráficos apresentam percentuais da quantidade de crimes, taxas de criminalidade em feriados e fins de semana, variação do número de crimes por dia e mês e, variação da criminalidade ao longo das horas do dia e ao longo da semana.

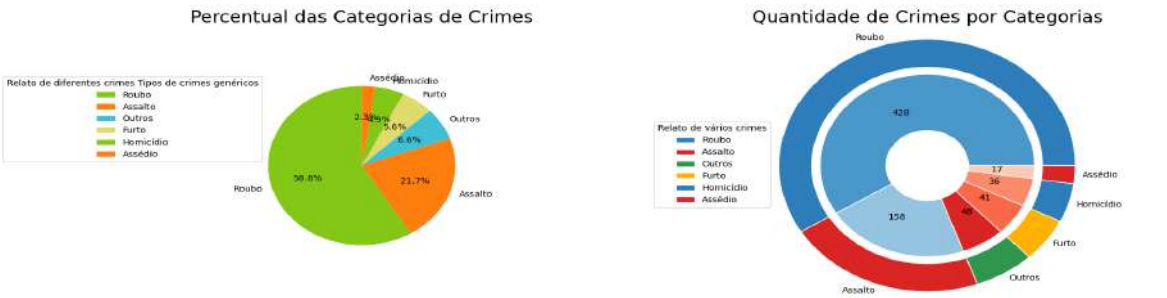
As figuras 6a e 6b apresentam o percentual das categorias de crimes, fornecendo uma visão sobre os diferentes tipos de crimes que estão incluídos tanto no conjunto de dados da Polícia como no Twitter. Nas figuras, também é possível observar a proporção de cada categoria de crime em ambas as bases de dados. No gráfico adjacente ao percentual de crimes é possível visualizar a quantidade de crimes por categoria em cada base. É

[13] https://github.com/mona2711/Data-Mining/tree/master

importante destacar que a quantidade de crimes do conjunto de dados da Polícia é substancialmente maior que em comparação ao do Twitter que possui uma quantidade maior de categorias de crimes.



(a) Base Policial



(b) Base Twitter

Figura 6. Percentual/Quantidade de crimes.

As figuras 7a e 7b demonstram a taxa de criminalidade em feriados, incluindo finais de semana. Esse tipo de informação é extremamente relevante para identificar quais são os momentos com maiores índices de criminalidade, se são durante a semana ou feriados e finais de semana. Nos gráficos é posssível notar os percentuais de crimes que acontecem em feriados e finais de semana. De acordo com os dados da base da Polícia, cerca de 70.2% dos crimes ocorrem durante a semana. Da mesma forma, no Twitter, aproximadamente 73.5% dos crimes também ocorrem nesse período. Esses números reforçam que a maioria dos crimes acontecem durante a semana.

(a) Base Policial **(b) Base Twitter**

Figura 7. Percentual de Criminalidade em Feriados e Finais de Semana.

Os gráficos 8a e 8b apresentam a variação do número de crimes por dia e mês em São Paulo. Em ambas as figuras é possível observar quais são os dias e meses mais perigosos ao longo do ano. Na base da Polícia, existem três datas com maiores índices de criminalidade. A primeira ocorreu nos dias 23 e 24 de fevereiro, no pré-carnaval, com um total de 2042 crimes registrados. A segunda foi no dia 3 de março, domingo de carnaval, com um total de 841 crimes registrados. A terceira data com maior índice de criminalidade foi em 23 de junho, durante a 23ª Parada LGBT de São Paulo, com um total de 1133 crimes registrados. É comum que nesses eventos ocorram grandes aglomerações de pessoas, o que pode criar uma ambiente propenso a crimes, como roubo, furtos e agressões. Por isso, no conjunto de dados da Polícia existem meses e dias que dobram a quantidade normal de boletins de ocorrências. Por outro lado, a base de dados do Twitter apresenta uma pequena quantidade de crimes relatados durante os anos de 2010 a 2021.

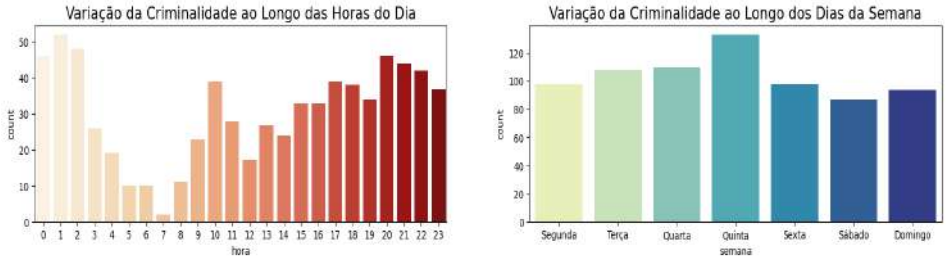
As figuras 9a e 9b apresentam a quantidade de ocorrências de crimes durante as 24 horas do dia para a polícia e para o Twitter. Para a base da polícia o pico de crimes ocorre entre as 17h e vai até às 23h. Já no Twitter o pico de crimes ocorre das 17h e vai até as 2h da manhã. Logo, é possível notar que a taxa de criminalidade em ambas as bases de dados aumenta a tarde, horário em que geralmente as pessoas estão saindo do trabalho. Não obstante, no gráfico adjacente é apresentado a quantidade de ocorrências de crimes durante os dias da semana, tanto para a Policia como para o Twitter. Analisando o gráfico, Terça e Quarta são os dias que mais acontecem crimes segundo a base da polícia. Já para o Twitter os dias da semana mais perigosos são Quarta e Quinta. Por outro lado, Domingo e Segunda são os dias com menos crimes na base da Polícia. No twitter Sábado e Domindo são os dias que menos possuem registros.

(a) Base Policial

(b) Base Twitter

Figura 8. Crimes Durante os Dias do Ano.

(a) Base Policial



(b) Base Twitter

Figura 9. Percentual/Quantidade de Crimes em Dias do Ano.

5.2. Modelos de Previsão

A fim de utilizar os dados da Polícia e do Twitter para construir modelos de classificação visando a previsão de crimes, serão utilizados dois classificadores: o Árvore de Decisão e o Naive Bayes. Serão consideradas informações como dia, mês, hora, bairro e se o dia é feriado/fim de semana ou não. Esses atributos serão utilizados para construir dois modelos de classificação, para cada conjunto de dados, tendo em vista verificar com estas carcterísticas o tipo de crime com base em dados relacionados à data e ao local dos eventos.

5.2.1. Classificador Naive Bayes

O classificador Naive Bayes é um algoritmo de aprendizado de máquina supervisionado eficaz e popular. Ele se basea no Teorema de Bayes (Fórmula 1) e é amplamente utilizado na área de ciência dos dados. Esse modelo estatístico é capaz de fazer previsões probabilísticas, o que o torna confiável para problemas de classificação em diferentes domínios.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{1}$$

Para a construção do modelo foi utilizada a biblioteca Scikit-Learn[14] do python que disponibiliza uma ampla gama de algoritmos e ferramentas para pré-processamento de dados, treinamento de modelos e mineração de dados. No modelo, foi utilizado o Naive Bayes, adequado para dados com distribuição gaussiana.

Os dados de crime utilizados no modelo possuem as seguinte caracteríticas: hora, dia, semana, bairro, feriado/fins de semana ou não. Para a classificação 25% dos dados

[14]https://scikit-learn.org/stable/

foram usados para teste, enquanto 75% foram usados para treinamento. O classificador foi treinado tanto para o conjunto de dados da Polícia, quanto para os dados do Twitter, para fazer uma previsão de determinados tipos de crimes contidos nos conjuntos de dados.

5.2.2. Classificador de Árvore de Decisão

O classificador de Árvore de Decisão é um algoritmo de aprendizado supervisionado usado para resolver problemas de classificação e regressão. Para a criação do modelo também foi utilizada a biblioteca Scikit-Learn do python. As mesmas carcterísticas do conjunto de dados aplicado no Naive Bayes também foram utilizados aqui. O modelo de Árvore de Decisão foi criado com base no critérios de entropia. A entropia é usada para escolher os atributos mais indicativos para a classificação. Para a classificação também foram utilizados 25% dos dados para testes e 75% para treino. A figura 10 apresenta a árvore construida para os dados do Twitter, mostrando que o atributo bairro foi selecionado como nó raiz para dividir os dados. A árvore apresentada tem no máximo 3 níveis de decisão.

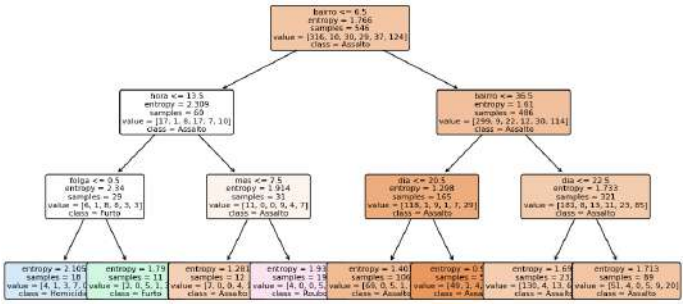


Figura 10. Árvore de Decisão dos dados do Twitter.

5.2.3. Análise

Para analisar qual classificador teve o melhor desempenho, foram avaliados os dois modelos construídos. Os modelos utilizados são: o Naive Bayes e a Árvore de Decisão que foram utilizados para prever o tipo de crime associado a uma combinação específica de hora, dia, mês, bairro e se é fim de semana ou não. O classificador Naive Bayes obteve uma acurância de 62% para o conjunto de dados do Twitter. Já para o conjunto de dados da Polícia o classificador obteve uma acurária de 52%. Por outro lado, o classificador de Árvore de Decisão obteve uma acurácia de 50% para o conjunto de dados da Polícia, enquanto que para o Twitter obteve 60% de acurácia.

Em seguida, os classificadores foram utilizados para prever os tipos de crimes em determinados locais. As caractéristicas necessárias para fazer a previsão são: hora, dia, mes, bairro e se o dia era final de semana ou não. Assim, era necessário que as variáveis setadas estivessem nos intervalos fornecidos pelas bases de dados. Ao realizar a simulação, é crucial utilizar como parâmetro somente os bairros que estão presentes nos 96 bairros utilizados para treinar o modelo com a base de dados da Polícia. Para o Twitter, por outro lado, apenas 89 bairros poderiam ser selecionados, uma vez que estes estão contidos no conjunto de dados disponível.

6. Conclusão e Trabalhos Futuros

Este trabalho apresentou um método utilizado para a extração de dados sobre criminalidade a partir de postagens realizadas na rede social Twitter, em português, para a cidade de São Paulo. Foram detalhadas as etapas realizadas até a construção de uma base de dados de 1.333 tweets geolocalizados e rotulados contendo relatos criminais. Os principais desafios encontrados estavam ligados a filtragem dos tweets a partir de palavras-chaves e a identificação das postagens por bairro. Algumas das limitações do trabalho são o uso da geolocalização dos tweets como aproximação para a localização dos crimes relatados pelos mesmos e a necessidade de se avaliar e rotular manualmente os tweets retornados pela consulta à API. Os dados do Twitter obtidos foram analisados junto a uma base de registros de boletins de ocorrência. A utilização do mesmo método apresentado neste trabalho, porém para outras cidades brasileiras, bem como o uso de técnicas de extração de localização a partir dos textos dos tweets podem ser considerados como duas opções de trabalhos futuros e promissores.

Além disso, nesse estudo, foram empregados dois algoritmos de classificação com o objetivo de prever tipos de crimes nos conjuntos de dados da Polícia e do Twitter. No conjunto de dados da Polícia, os classificadores Naive Bayes e Árvore de Decisão obtiveram sequencialmente acurácias de 52% e 50%, respectivamente. Já no conjunto de dados do Twitter, os classificadores alcançaram acurácias sequenciais de 62% e 60%. Utilizando esses classificadores é possível obter insights sobre o comportamento dos crimes em distintos locais e horários da cidade, contribuindo para um melhor entendimento dos padrões e identificando quais são os tipos de crimes mais recorrentes. A utilização desses classificadores desempenha um papel fundamental na análise dos dados criminais, oferecendo uma visão mais ampla e facilitando o entendimento dos padrões e comportamentos de dados criminais.

Referências

Adesola, F., Misra, S., Omoregbe, N., Damasevicius, R., and Maskeliunas, R. (2019). *An IOT-Based Architecture for Crime Management in Nigeria*, pages 245–254. Springer Singapore, Singapore.

Almanie, T., Mirza, R., and Lor, E. (2015). Crime prediction based on crime types and using spatial and temporal criminal hotspots. *arXiv preprint arXiv:1508.02050*.

Cai, L. and Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Sci. J.*, 14:2.

Clarindo, J. P., Coutinho, F., and Freitas, A. L. (2016). Detecção de casos de violência patrimonial a partir do twitter. In *Anais do V Brazilian Workshop on Social Network Analysis and Mining*, pages 211–216. SBC.

dos Reis, G. O. and Nakamura, E. F. (2017). Crimes: reportes oficiais vs. postagens no twitter. In *Anais Estendidos do XXIII Simpósio Brasileiro de Sistemas Multimídia e Web*, pages 111–114. SBC.

dos Santos, L. S. F. C. (2015). Estudo online da dinâmica espaço-temporal de crimes através de dados da rede social twitter. Master's thesis, Universidade Federal de Minas Gerais, Belo Horizonte.

Gerber, M. S. (2014). Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, 61:115–125.

Khan, M., Ali, A., and Alharbi, Y. (2022). Predicting and preventing crime: A crime prediction model using san francisco crime data by classification techniques. *Complexity*, 2022.

Laufs, J., Borrion, H., and Bradford, B. (2020). Security and the smart city: A systematic review. *Sustainable Cities and Society*, 55:102023.

Mahajan, R. and Mansotra, V. (2021). Correlating crime and social media: using semantic sentiment analysis. *International Journal of Advanced Computer Science and Applications*, 12(3).

Neto, A. J. V., Zhao, Z., Rodrigues, J. J. P. C., Camboim, H. B., and Braun, T. (2018). Fog-based crime-assistance in smart iot transportation system. *IEEE Access*, 6:11101–11111.

Prathap, B. R. and Ramesha, K. (2018). Twitter sentiment for analysing different types of crimes. In *2018 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, pages 483–488. IEEE.

Sarhan, Q. I. (2020). Systematic survey on smart home safety and security systems using the arduino platform. *IEEE Access*, 8:128362–128384.

Secron, T. M., da Silva, E. R., de Farias, C. M., and Cruz, T. (2016). Sigaciente: Uma ferramenta para inferência do trânsito e de rotas seguras baseada em dados sociais. In *ERSI'2016*, pages 58–65.

Vivek, M. and Prathap, B. R. (2023). Spatio-temporal crime analysis and forecasting on twitter data using machine learning algorithms. *SN Computer Science*, 4(4):383.