DataCrimeBR: Construção de um Dataset de Crimes Reportados em Tweets no Brasil

Miguel A. R. e Silva[1], **Philipe de Freitas Melo**[1], **Thais R. M. Braga Silva**[1]

[1]Instituto de Ciências Exatas e Tecnológicas (IEF)
Universidade Federal de Viçosa (UFV) - Florestal – MG – Brasil

{miguel.a.silva, philipe.freitas, thais.braga}@ufv.br

Resumo. *Redes sociais como o Twitter/X são amplamente usadas para compartilhar experiências e acontecimentos, incluindo relatos de crimes. A identificação automática desses relatos enfrenta desafios, especialmente pela falta de dados em português que capturem a ambiguidade linguística e o uso informal da linguagem, dificultando a distinção entre descrições factuais e expressões figurativas. O DataCrimeBR reúne 61.715 tweets em português, obtidos por um processo rigoroso de curadoria, categorização de tipos de crimes e aplicação de filtros para garantir qualidade linguística e temática. O conjunto foi enriquecido com análises de sentimento, detecção de toxicidade e reconhecimento de entidades de localização geográfica, oferecendo um recurso robusto para pesquisas em Processamento de Linguagem Natural e segurança pública, útil no desenvolvimento e avaliação de sistemas de detecção de relatos criminais em ambientes digitais.*

Abstract. *Social networks such as Twitter/X are widely used to share experiences and events, including crime reports. The automatic identification of these reports faces challenges, especially due to the lack of Portuguese-language datasets that capture linguistic ambiguity and informal language use, which hinder the distinction between factual descriptions and figurative expressions. The DataCrimeBR dataset comprises 61,715 tweets in Portuguese, obtained through a rigorous curation process, crime-type categorization, and the application of filters to ensure linguistic and thematic quality. The dataset was also enriched with sentiment analysis, toxicity detection, and geographic location entity recognition, offering a robust resource for research in Natural Language Processing and public safety, useful for developing and evaluating systems aimed at detecting crime reports in digital environments.*

1. Introdução

As redes sociais revolucionaram a maneira como as pessoas interagem e compartilham informações na Web. Com o advento dessas plataformas, a comunicação se tornou mais rápida, dinâmica e acessível, rompendo barreiras geográficas e culturais. Entre estas plataformas, destaca-se o Twitter/X, com mais de 500 milhões de tweets diários [Dunn 2024], sendo umas das redes mais utilizadas para a disseminação de todo tipo de informação, desde notícias em tempo real até movimentos sociais e culturais.

Entre os conteúdos divulgados no Twitter/X, destacam-se os relatos de crimes, nos quais usuários reportam furtos, assédios, roubos ou assaltos que sofreram ou

testemunharam. No Brasil, onde a violência é alta e a confiança nos canais formais de denúncia pode ser limitada, as redes sociais permitem mapear incidentes em tempo real e em larga escala, inclusive casos que não chegam às autoridades, já que mais da metade das vítimas de roubo e furto não registra boletim de ocorrência [Abdala 2022].

Entretanto, a identificação de relatos de crimes em redes sociais não é trivial e apresenta desafios, especialmente no português, cuja complexidade gramatical, diversidade de expressões regionais e uso frequente de gírias e linguagem informal dificultam a detecção automática. Termos como "roubou" ou "assaltou" podem aparecer em sentidos figurativos ou humorísticos, confundindo modelos mais simples. Esses fatores exigem abordagens que considerem não apenas palavras-chave, mas também nuances culturais e linguísticas locais. Desta forma, apresentamos o ***DataCrimeBR***, uma base de dados robusta e enriquecida para aplicação em sistemas automatizados de detecção de crimes. Além disso, aprofundamos a investigação desse contexto específico de ambiguidade, no qual abordagens mais genéricas ou baseadas em léxicos costumam falhar. Para isso, construímos um *dataset* com 61.715 tweets rotulados, a partir de uma extensa coleta de dados online, enriquecidos com análises de sentimento, detecção de toxicidade e reconhecimento de entidades nomeadas (NER), com ênfase na localização geográfica.

O restante do trabalho se divide na seguinte estrutura: Na Seção 2, discutimos trabalhos relacionados que abordam o uso de redes sociais para identificação de crimes. Em seguida, na Seção 3, detalhamos a construção do *dataset*. Por fim, na Seção 4, apresentamos a conclusão e sugestões para trabalhos futuros.

2. Trabalhos Relacionados

Dados de redes sociais, especialmente do Twitter/X, oferecem ampla fonte de conhecimento online e têm sido aplicados em diversos contextos, como análise de sentimento [Giachanou and Crestani 2016], extração de informações geográficas [Zheng et al. 2018] e detecção de eventos em tempo real [Vieweg et al. 2010, Earl et al. 2013]. Também são usados em estudos de saúde, como avaliação de bem-estar mental [Coppersmith et al. 2014] e identificação de depressão pós-parto [De Choudhury et al. 2014].

No contexto de segurança pública, a identificação de crimes reais em postagens online é relevante, pois muitos não são formalmente relatados [Abdala 2022]. Trabalhos anteriores coletaram e classificaram mensagens relacionadas a crimes [Abbass et al. 2020, Lombo et al. 2022, Shoeibi et al. 2021], explorando diferentes *features*, como dados do *Sentiment140* [Go et al. 2009] adaptados à detecção de crimes [Bokolo et al. 2024], combinações de informações temporais, geográficas e sociais [Sandagiri et al. 2020a, Sandagiri et al. 2020b], e abordagens linguísticas como *part-of-speech tagging*, *Brown clustering* [Vo et al. 2020] e *ensemble* com *TF-IDF* [Siddiqui et al. 2023].

No Brasil, estudos coletaram e analisaram dados criminais em português [Clarindo et al. 2016], mapeando áreas afetadas [Almeida 2018], aplicando NER e geolocalização [Patricio 2023], ou construindo bases para análises estatísticas e classificação [dos Santos 2015].

Embora estes estudos apresentem avanços importantes para a identificação de crimes em ambientes online, poucos se aprofundam na análise das ambiguidades e peculiari-

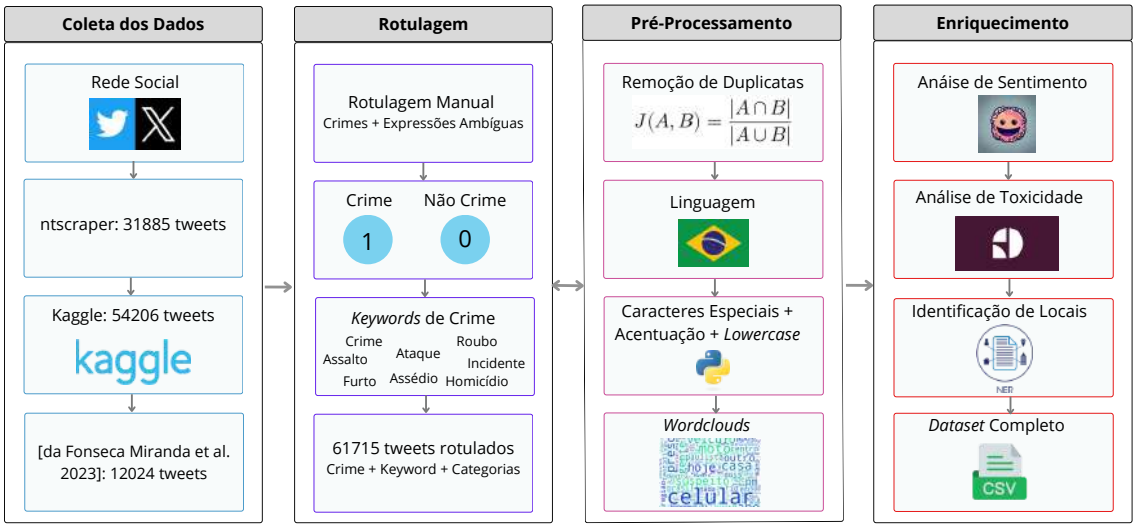


Figura 1. Diagrama das etapas de construção do *dataset* do DataCrimeBR.

dades linguísticas que afetam diretamente a qualidade e a efetividade dos dados utilizados nestas aplicações. Expressões como "roubou meu coração" ilustram as dificuldades que algoritmos enfrentam ao lidar com múltiplos sentidos e nuances contextuais. Este trabalho é complementar aos outros ao buscar preencher essa lacuna ao construir um *dataset* focado especificamente em casos ambíguos, aprofundando no entendimento desse problema e preparando uma base sólida para o desenvolvimento de modelos mais sensíveis às complexidades da linguagem do português brasileiro.

3. Construção de um *dataset* sobre relatos criminais

A construção de um *dataset* robusto e abrangente voltado para relatos criminais é uma etapa essencial para garantir a qualidade e efetividade de modelos de aprendizado de máquina supervisionados. Nesta seção, organizamos o processo para construção da base de dados proposta por este trabalho, dividido em quatro etapas distintas: (i) Coleta dos Dados, (ii) Rotulagem, (iii) Pré-Processamento e (iv) Enriquecimento do *Dataset*, conforme ilustrado na Figura1, que apresenta o fluxo completo desde a aquisição dos dados até o refinamento e enriquecimento final. Todo o processo foi implementado integralmente em *Python*[1], utilizando bibliotecas específicas para cada etapa. Em seguida, entraremos em detalhes de cada fase do processo de construção.

3.1. Coleta dos dados

Devido a limitações recentes impostas pelo X, o acesso à API[2] da plataforma foi restrito [Barnes 2023], tornando a tarefa de aquisição de dados ainda mais complexa atualmente e gerando a necessidade de buscar alternativas para se estudar a plataforma. Para contornar este desafio, focamos nossa coleta em **três fontes principais de dados**: (i) uso de um *WebScraper* (ii) uma base de dados pública obtida via Kaggle, e (iii) uma base de dados de crimes já compilada por um trabalho anterior de [da Fonseca Miranda et al. 2023], que confirmou forte similaridade entre o padrão de crimes extraídos de tweets para a cidade de São Paulo e os registros oficiais da polícia,

[1] https://www.python.org/
[2] https://developer.x.com/en/products/x-api

demonstrando que a base é uma boa referência para representar ocorrências criminais. Como resultado final, obtivemos 61.715 tweets rotulados de maneira sistemática com o foco na extração de informações sobre crimes.

Para a primeira etapa do *WebScraper*, foi desenvolvido um *script* utilizando a biblioteca *ntscraper*[3] para coletar tweets com base em palavras-chave previamente definidas. Essas palavras englobavam termos associados a crimes comuns em ambientes urbanos, baseados no trabalho de [dos Santos 2015], como "assalto", "roubo", "furto", "assédio", "crime", "ataque", "incidente"e "homicídio", além de novos termos correlatos e variações de gênero, número e grau. Considerando que, no Twitter/X, usuários frequentemente mencionam locais das ocorrências, foram incluídos também termos como "rua", "praça", "avenida"e "bairro", a fim de direcionar a coleta para mensagens mais relevantes e reduzir ruídos. Ao todo, aplicaram-se mais de 100 variações de busca. A coleta foi realizada no início de janeiro de 2024, de forma retroativa, abrangendo todo o histórico de tweets até aquela data e resultando em 31.885 mensagens. Cabe destacar que, nesse período, o Twitter já havia encerrado sua API oficial [Hernandes 2023], o que inviabilizou coletas em larga escala via métodos tradicionais. Por isso, recorreu-se a um *WebScraper* personalizado, permitindo obter, de forma eficiente, um volume satisfatório de dados para este estudo.

A segunda fonte de dados foi obtida por meio de um *dataset* disponibilizado no *Kaggle*[4], extraindo um total de 54.206 tweets gerais em português utilizados para nosso estudo. Com o objetivo de equilibrar melhor o *dataset* e construir um conjunto mais robusto, as mensagens do *Kaggle* foram filtradas da seguinte maneira: 2.491 tweets contendo apenas palavras-chave relacionadas a crimes; 1.715 tweets contendo exclusivamente palavras-chave de locais, sem necessariamente se referir a crimes e os 50.000 tweets restantes foram amostrados aleatoriamente sobre tópicos gerais e assuntos diversos não relacionados ao contexto de crime.

Por fim, a terceira base de dados foi obtida diretamente a partir do trabalho de [da Fonseca Miranda et al. 2023], que disponibilizaram um *dataset* contendo 12.024 tweets coletados entre 2010 e 2022 da cidade de São Paulo. Esses tweets foram extraídos via API do Twitter/X e focam em relatos de crimes na cidade. Com base nas três fontes de dados, construímos um corpus inicial com 98.115 tweets. Em seguida, esse material foi submetido a um processo sistemático de rotulação, conduzido por um único pesquisador responsável pelo estudo, resultando na versão final do conjunto de dados.

3.2. Rotulagem e taxonomia dos dados

No contexto deste trabalho, rotulamos as mensagens do nosso corpus, utilizando a biblioteca *pandas*[5], de forma binária entre **Crime** e **Não-Crime**, dando destaque também para aquelas instâncias que possuíam algum termo relacionado ao contexto criminal, mas não representavam um crime de fato.

De forma mais específica, dois níveis de rotulação foram utilizados neste processo: o primeiro – binário – sobre ***Crime e Não-Crime***: Indica se um tweet realmente relata um

[3] https://pypi.org/project/ntscraper/

[4] https://www.kaggle.com/datasets/augustop/portuguese-tweets-for-sentiment-analysis/data

[5] https://pandas.pydata.org/

crime, sendo classificado como 1 se sim e 0 se não na coluna de Crime do dataset disponibilizado. O segundo níve é o referente à coluna ***Keyword Criminal*** que indica se um tweet contém palavras-chave relacionadas a crimes, como "assalto", "roubo", "furto", etc. A classificação é 1 se contém essas palavras-chave e 0 caso contrário. Este rótulo último é importante para diferenciarmos as expressões ambíguas que possuem rótulo de *Não-Crime*, mas que contêm palavras-chave no contexto criminal. Para fins de entendimento, referimos a esse conjunto específico como ***Expressões Ambíguas***.

Além disso, como terceira etapa, para os tweets classificados como crimes, rotulamos manualmente também a categoria específica de cada um. Vale observar que este é um problema multi-rótulo, ou seja, um tweet pode ser classificado em mais de uma categoria, caso haja múltiplas referências ou tipos de crimes mencionados. As categorias foram definidas da seguinte maneira: ***Furto***: Refere-se a subtração de bens sem qualquer tipo de violência ou ameaça direta à vítima. ***Roubo***: Consiste na subtração de bens com o uso de violência ou ameaça grave. ***Assalto***: O termo é frequentemente usado de forma genérica para descrever roubos. [de Araujo 2017]. ***Assédio***: Engloba relatos de qualquer forma de importunação, agressão, assédio, exploração sexual ou estupro [Beserra 2022]. ***Homicídio***: Refere-se ao ato de matar alguém, independentemente das circunstâncias, incluindo homicídios dolosos e culposos [Ciardo 2015]. ***Segurança Pública***: Tweets com tema sobre crimes, mas não indicam que o usuário tenha sido vítima ou tenha presenciado tal crime. Isso inclui alertas gerais sobre segurança, descrições de áreas de risco ou discussões sobre crimes passados, muitas vezes visando alertar [Vedova 2018]. ***Outros***: Tweets que não se encaixam nas categorias anteriores, mas que ainda estão relacionadas a incidentes criminais, ataques ou comportamentos criminosos de alguma forma.

Essas categorias foram levantadas a partir de uma anotação empírica parcial dos dados e de forma iterativa, refinando os rótulos a cada etapa até chegar neste conjunto final de categorias. Para esta etapa, o dataset conta com uma coluna individual referente a cada um dos rótulos sendo 1 caso o tweet relate aquele crime e 0 caso contrário. Ao analisar os dados, primeiramente, optamos por rotular manualmente 10.000 tweets que possuíam palavras-chave de crime, incluindo todos os 2.491 tweets contendo apenas palavras-chave relacionadas a crimes provenientes do *Kaggle* mais uma amostra de 7.509 instâncias coletadas via *ntscraper* e da base de [da Fonseca Miranda et al. 2023] para manter a amostra balanceada, totalizando 10.000 tweets rotulados. Destes, 5.000 tweets foram rotulados como *Crime*, e os outros 5.000 como *Não-Crime* e *Keyword Criminal = 1*, ou seja, como *Expressões Ambíguas*.

Por fim, os 51.715 tweets restantes provenientes do Kaggle, sobre tópicos gerais e que não relatam crimes, foram rotulados automaticamente como ***Não-Crime*** pois não contêm informações sobre crimes nem palavras-chave relacionadas.

3.3. Pré-processamento do texto

No pré-processamento do texto, aplicamos várias estratégias para melhorar a qualidade e a consistência do nosso corpus. Primeiro, utilizamos o índice de *Jaccard*, implementado através da biblioteca *scikit-learn*[6], para excluir duplicatas. Também empregamos a biblioteca *langdetect*[7] para remover qualquer tweet em outro idioma que não português. Por

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

[7]https://pypi.org/project/langdetect/

fim, realizamos a limpeza do texto utilizando *re*[8] para remover *links*, menções, caracteres especiais e acentos, além de converter todas as letras para minúsculas. Para ilustrar os dados, a Figura 2 mostra as nuvens de termos de cada categoria de crime rotulada, geradas com a biblioteca *wordcloud*[9].

Na categoria ***Assédio***, as mensagens apresentam tom claro e direto, com destaque para termos ligados a crimes sexuais, locais como "escola" e "rua", e referências ao público feminino, como "mulher" e "menina". Diferente de outras categorias, "polícia" não aparece com destaque. Em ***Roubo*** e ***Furto***, termos como "celular", "moto" e "carro" indicam que dispositivos móveis e automóveis estão entre os principais alvos. Na categoria ***Assalto***, a presença de "casa" sugere relação com subtração de propriedades residenciais. Já em ***Segurança Pública***, a palavra "medo" evidencia a apreensão da população quanto à segurança. Por fim, "ônibus" aparece em algumas categorias, indicando seu uso frequente como local de crimes.

(a) **Roubo** (b) **Assalto** (c) **Furto**

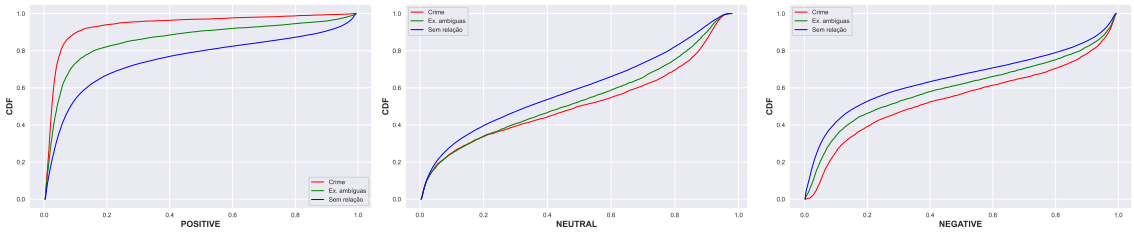
(d) **Homicídio** (e) **Assédio** (f) **Segurança Pública**

Figura 2. Nuvens de Palavras

3.4. Enriquecimento do *Dataset*

Para tornar o *dataset* mais robusto e informativo, incorporamos três tipos de análises. Primeiro, realizamos uma **análise de sentimento**, identificando se os tweets apresentam tendência positiva, negativa ou neutra, o que auxilia a compreender o tom predominante

[8] https://docs.python.org/3/library/re.html
[9] https://amueller.github.io/word_cloud/



(a) **Positivo** (b) **Neutro** (c) **Negativo**

Figura 3. CDFs Análise de Sentimento

nas publicações. Em seguida, aplicamos uma **análise de toxicidade**, classificando os tweets em categorias como discurso tóxico, ameaças, insultos, ataques pessoais, xingamentos e outros comportamentos agressivos, com o intuito de identificar conteúdos potencialmente nocivos e entender sua relação com o contexto criminal. Por fim, implementamos o **reconhecimento de entidades nomeadas** focado em locais, já que observamos que tweets que relatam crimes frequentemente mencionam regiões ou pontos específicos onde os eventos ocorreram.

3.4.1. Análise de Sentimento

A análise de sentimento é essencial para captar as emoções e opiniões expressas em um texto. Essa classificação permite identificar a carga emocional associada aos relatos de crimes, facilitando a compreensão do impacto psicológico e social dos eventos descritos.

Para enriquecer a base de dados, realizamos a análise de sentimento em todos os tweets usando o modelo *PySentimiento* [Pérez et al. 2021], classificando os textos entre positivo, negativo ou neutro. Esse modelo fornece uma pontuação de sentimento para cada polaridade em formato percentual. Por exemplo, para o tweet "*assaltaram o menino pertinho da minha rua*", o modelo atribuiu as pontuações de 0,82% positivo, 1,35% neutro e 97,84% negativo, indicando uma carga emocional fortemente negativa.

Com essa etapa, realizamos uma análise quantitativa detalhada da relação entre o sentimento do texto e os crimes. Primeiro, observamos a CDF (*Cumulative Distribution Function*) para visualizar a distribuição de cada polaridade (positividade, neutralidade, negatividade) nas diferentes categorias de tweets, conforme mostrado na Figura 3. Além disso, aplicamos o teste de *Kolmogorov-Smirnov* para avaliar estatisticamente as diferenças entre as distribuições das categorias. Os resultados indicam que tweets de *Crime* e *Não-Crime* apresentam distribuições estatisticamente diferentes entre si.

Observamos que tweets classificados como ***Crime*** apresentam baixa positividade, alta negatividade e pouca neutralidade, destacando-se das demais categorias. Já os tweets ***Não-Crime*** tendem a ser mais positivos e neutros, com baixa negatividade, evidenciando um distanciamento claro do contexto criminal. Também destacamos separadamente os tweets com ***Expressões Ambíguas***, ou seja, aqueles com termos sobre crime, mas categorizados como Não-Crime. Estes exibem uma distribuição intermediária, mas ainda mantendo alta negatividade. Esses resultados sugerem que, embora não sejam suficiente para distinguir as classes, o sentimento pode ser uma *feature* útil na classificação dessas mensagens.

É interessante notar que alguns relatos de crime apresentam sentimento positivo, como expressões de solidariedade, agradecimentos às autoridades ou alívio em situações de quase-crime. Já os relatos neutros tendem a ser informativos e factuais, com pouca carga emocional. Além disso, a análise de sentimento auxilia na interpretação de mensagens ***Não-Crime*** que contêm palavras associadas a crimes, nas quais sentimentos negativos podem indicar frustração ou ironia, enquanto sentimentos positivos podem ocorrer em contextos afetivos ou de entretenimento.

3.4.2. Análise de Toxicidade

Detectar toxicidade em textos ajuda a identificar linguagem agressiva, desrespeitosa e até mesmo ameaçadora, especialmente em ambientes de redes sociais, no qual o discurso tóxico pode afastar usuários e impactar discussões. Para isso, utilizamos a *Perspective-API*[10], acessada por meio da biblioteca *googleapiclient*[11], em nosso *dataset* para medir diferentes tipos de toxicidade em tweets.

Esta ferramenta fornece modelos pre-treinados com aprendizado profundo para várias categorias de toxicidade, que ajudam a compreender melhor as nuances da linguagem em textos potencialmente ofensivos em vários idiomas, inclusive com modelos treinados em português. Embora o principal modelo seja usado para medir o nível de *Toxicidade* de um texto, a API também oferece ferramentas para analisar *Insultos*, *Ameaças*, *Xingamentos*, *Ataques a Identidade*, *Toxicidade Severa*, que avaliam diferentes aspectos de um conteúdo ofensivo.

Cada métrica varia de 0 a 1, sendo que valores próximos a 1 indicam maior probabilidade de um comentário se enquadrar na categoria. Para nosso estudo, consideramos, após alguns testes empíricos com várias faixas de valores, que caso o texto resultasse num valor maior que $0,6$, que ele então se enquadra nesta categoria de toxicidade para nosso contexto. Em um exemplo, o tweet "*dois assedios na rua em menos de 5 minutos p**** vai se f****" teve um resultado de $0,9446$ de *Toxicidade*, portanto, é considerado um comentário tóxico.

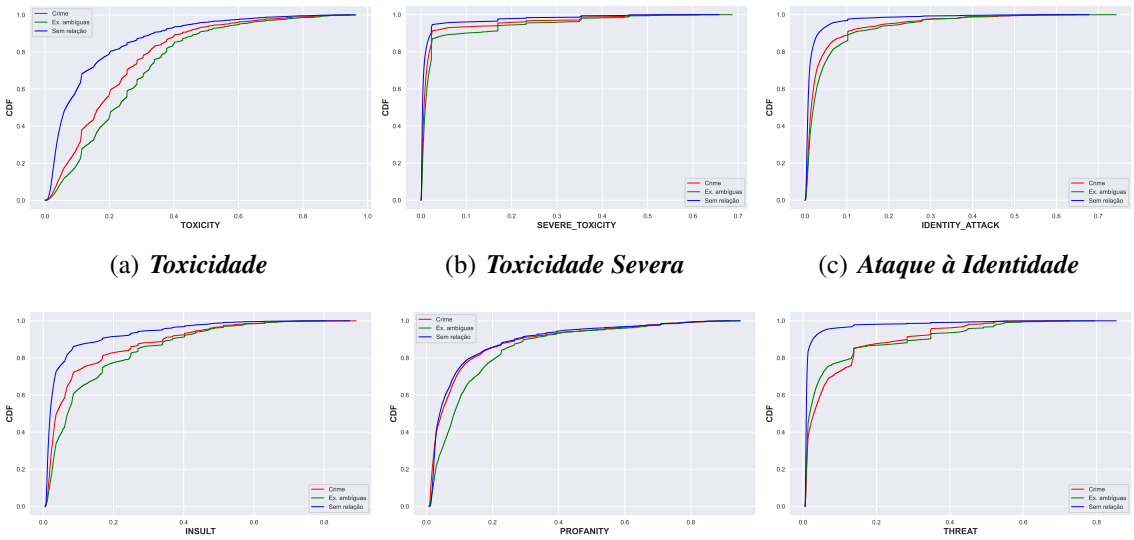
Entre os 61.715 tweets analisados, as categorias apresentam as seguintes proporções: *Toxicidade* corresponde a 2,83% do total, *Xingamentos* a 3,26%, *Insulto* a 0,6%, *Ameaça* a 0,14%, enquanto *Toxicidade Severa* e *Ataque à Identidade* representam cada uma apenas 0,04% dos tweets. A Figura 4 apresenta as distribuições acumuladas (CDF) dos resultados de toxicidade e suas subcategorias ao longo das diferentes classes de tweets. Observa-se que as categorias ***Crime*** e ***Expressões Ambíguas*** concentram escores mais elevados de toxicidade, insulto e ameaça em comparação com a classe ***Não-Crime***. Para validar estatisticamente essas diferenças, novamente, foi aplicado o teste de *Kolmogorov-Smirnov*, indicando divergências significativas entre as distribuições e confirmando a associação entre conteúdo criminal e maior presença de linguagem tóxica nos textos.

Os tweets da classe ***Crime*** oferecem uma visão clara de suas características. Os resultados de *Toxicidade* e *Toxicidade Severa* são marcadas por raiva intensa e frustração como em "*são paulo ta f*** muito assalto e mortes*". Os resultados de *Ataque de Identi-*

[10] `https://perspectiveapi.com/`

[11] `https://github.com/googleapis/google-api-python-client`

dade envolvem geralmente estereótipos, como na mensagem "*ser assediada pelos pretos do meu bairro mas que típico kkkk*", enquanto *Ameaça* são mensagens diretas e muitas vezes motivadas por justiça própria, como "*mt triste que assaltaram minha mãe cacete mano espero que tome no c* e exploda esse escr****".



(a) ***Toxicidade*** (b) ***Toxicidade Severa*** (c) ***Ataque à Identidade***

(d) ***Insulto*** (e) ***Xingamentos*** (f) ***Ameaça***

Figura 4. CDFs Análise de Toxicidade

3.4.3. Identificação de locais

Saber se um texto se refere a um local geográfico é frequentemente apontado na literatura como uma característica relevante para identificação de crimes [Patricio 2023, Sandagiri et al. 2020a, Sandagiri et al. 2020b]. Portanto, como última fase do enriquecimento, adicionamos um atributo se o tweet se refere a um local ou não.

Para realizar essa identificação, utilizamos o modelo *large* de [Guillou 2021] para identificação de entidades nomeadas, carregado por meio da biblioteca *transformers*[12], que apresentou os melhores resultados para o português em relação a outras opções testadas, incluindo ferramentas disponibilizadas por [Tedeschi et al. 2021, Guillou 2021] *base*, além da biblioteca spaCy[13]. Após essa análise comparativa, [Guillou 2021] *large* demonstrou-se o mais adequado para nossa aplicação.

Executamos o modelo em todos os dados da nossa base, mantendo exclusivamente a entidade "local". Caso um local fosse identificado, atualizamos o *label* ***Local*** para o valor 1. Com esse processo, detectamos que **3456** tweets (5,6% do total) do nosso *dataset* possuem locais especificados. Dos que possuem local, **1874** são crimes e **1582** não são. Esse resultado nos permite observar uma forte correlação entre a presença de um local nos relatos e a menção de crimes, indicando que, além de descrever o evento, as pessoas frequentemente incluem informações sobre o local onde ele ocorreu. Essa característica foi posteriormente avaliada para verificar se a presença dessa informação influenciaria o desempenho dos modelos na identificação de tweets com relatos criminais.

[12]`https://huggingface.co/docs/transformers/index`
[13]`https://spacy.io/`

3.5. Desafios e Limitações

Apesar dos esforços para assegurar qualidade e representatividade, o *DataCrimeBR* apresenta limitações inerentes a dados de redes sociais. O uso de palavras-chave na coleta pode gerar viés temático, e o contexto informal e ambíguo dos tweets dificulta a rotulagem, especialmente em casos de ironia, sarcasmo ou regionalismos.

Além disso, por se restringir ao Twitter/X, há limitações na generalização dos resultados para outras plataformas ou formas de comunicação online além de não garantir uma representatividade da demografia brasileira. Esses fatores devem ser considerados por pesquisadores que pretendem utilizar a base em aplicações supervisionadas ou análises sensíveis à semântica. Entretanto, no país, mais da metade das vítimas de roubo e furto não registra boletim de ocorrência [Abdala 2022], o que torna outros dados estatísticos sobre o assunto também potencialmente defasados. Portanto, com 23 milhões de usuários [WPR 2025] no Brasil e por seu caráter mais diário, esse dataset pode ser uma fonte complementar desta informação sobre criminalidade no país.

4. Conclusão e Trabalhos Futuros

Este artigo apresentou a construção e disponibilização do *DataCrimeBR*, disponível em `https://zenodo.org/records/15724169`, um conjunto de dados composto por 61.715 tweets em língua portuguesa, rotulados quanto à presença ou não de relatos criminais. A base foi enriquecida com informações adicionais, incluindo análises de sentimento, toxicidade e identificação de locais mencionados, a fim de fornecer contexto semântico e estrutural para estudos futuros em Processamento de Linguagem Natural (PLN), segurança pública e análise social. O dicionário do *dataset* é composto por atributos que incluem o identificador único e o texto pré-processado do tweet, métricas de toxicidade e sentimento (valores contínuos de 0 a 1) e rótulos binários (0 ou 1) que indicam tipos específicos de crime, presença de palavra-chave relacionada e menção a local.

Devido às restrições de acesso à API do Twitter/X, a obtenção desse tipo de dado tornou-se cada vez mais rara. O *DataCrimeBR* se destaca por tratar a ambiguidade linguística de expressões sobre crimes em português, frequentemente usadas em contextos figurativos ou não criminais, por meio de um processo criterioso e manual de rotulação. O conjunto oferece um recurso robusto para detecção de crimes em dados textuais, útil tanto no desenvolvimento de modelos de IA quanto na avaliação de abordagens sensíveis à linguagem informal.

Como trabalhos futuros, pretende-se ampliar a coleta para outras plataformas, incluir novos tipos de crimes e explorar seu uso no *benchmarking* de modelos de classificação voltados à detecção de eventos, análise de discurso e identificação de linguagem violenta ou discriminatória.

Referências

[Abbass et al. 2020] Abbass, Z., Ali, Z., Ali, M., Akbar, B., and Saleem, A. (2020). A framework to predict social crime through twitter tweets by using machine learning. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, pages 363–368.

[Abdala 2022] Abdala, V. (2022). Pesquisa do IBGE mostra subnotificação de roubos e furtos no Brasil. Online. Agência Brasil.

[Almeida 2018] Almeida, T. L. M. d. (2018). Estudo sobre aplicação de aprendizado de máquina para identificação de assaltos através de informações do twitter.

[Barnes 2023] Barnes, J. (2023). Twitter Ends Its Free API: Here's Who Will Be Affected — forbes.com. https://www.forbes.com/sites/jenaebarnes/2023/02/03/twitter-ends-its-free-api-heres-who-will-be-affected/.

[Beserra 2022] Beserra, T. (2022). Quais são os tipos de crimes sexuais previstos no Brasil? https://www.jusbrasil.com.br/artigos/quais-sao-os-tipos-de-crimes-sexuais-previstos-no-brasil/1642386524. [Accessed 09-01-2025].

[Bokolo et al. 2024] Bokolo, B. G., Onyehanere, P., Ogegbene-Ise, E., Olufemi, I., and Tettey, J. N. A. (2024). Leveraging machine learning for crime intent detection in social media posts. In Zhao, F. and Miao, D., editors, *AI-generated Content*, pages 224–236, Singapore. Springer.

[Ciardo 2015] Ciardo, F. (2015). Do Homicídio - Artigo 121 do Código Penal. https://www.jusbrasil.com.br/artigos/do-homicidio-artigo-121-do-codigo-penal/177410501. [Accessed 09-01-2025].

[Clarindo et al. 2016] Clarindo, J., Coutinho, F., and Freitas, A. (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, Porto Alegre, RS, Brasil. SBC.

[Coppersmith et al. 2014] Coppersmith, G., Dredze, M., and Harman, C. (2014). Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60.

[da Fonseca Miranda et al. 2023] da Fonseca Miranda, G. V., Almeida, V. G. d. J., Silva, T. R. B., and Silva, F. A. (2023). Extraçao e avaliaçao de uma base de dados sobre criminalidade em português a partir do twitter. In *Anais do XV Simpósio Brasileiro de Computação Ubíqua e Pervasiva*, pages 61–70. SBC.

[de Araujo 2017] de Araujo, A. A. (2017). Qual a diferença entre furto e roubo? https://www.jusbrasil.com.br/artigos/qual-a-diferenca-entre-furto-e-roubo/447365236. [Accessed 09-01-2025].

[De Choudhury et al. 2014] De Choudhury, M., Counts, S., Horvitz, E. J., and Hoff, A. (2014). Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 626–638.

[dos Santos 2015] dos Santos, L. S. F. C. (2015). Estudo online da dinâmica espaço-temporal de crimes através de dados da rede social twitter.

[Dunn 2024] Dunn, N. (2024). Top 26 X (Formerly Twitter) Statistics. Online. Charle Agency.

[Earl et al. 2013] Earl, J., McKee Hurwitz, H., Mejia Mesinas, A., Tolan, M., and Arlotti, A. (2013). This protest will be tweeted: Twitter and protest policing during the pittsburgh g20. *Information, communication & society*, 16(4):459–478.

[Giachanou and Crestani 2016] Giachanou, A. and Crestani, F. (2016). Like it or not: A survey of twitter sentiment analysis methods. *ACM Comput. Surv.*, 49(2).

[Go et al. 2009] Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

[Guillou 2021] Guillou, P. (2021). Nlp: Modelos e web app para reconhecimento de entidade nomeada (ner) no domínio jurídico. Acesso em: 19 nov. 2024.

[Hernandes 2023] Hernandes, R. (2023). Mudança no Twitter cria dificuldade para pesquisadores com extração e análise de dados. Folha de São Paulo. https://www1.folha.uol.com.br/poder/2023/02/mudanca-no-twitter-dificulta-pesquisadores-com-extracao-e-analise-de-dados.shtml. [Accessed 08-08-2025].

[Lombo et al. 2022] Lombo, X., Oyelade, O. N., and Ezugwu, A. E. (2022). Crime detection and analysis from social media messages using machine learning and natural language processing technique. In Gervasi, O., Murgante, B., Misra, S., Rocha, A. M. A. C., and Garau, C., editors, *Computational Science and Its Applications 2022 Workshops*, pages 502–517, Cham. Springer.

[Patricio 2023] Patricio, G. S. (2023). Criação de um aplicativo para mapeamento da criminalidade da cidade de belo horizonte por meio de atividade crowdsourcing no twitter.

[Pérez et al. 2021] Pérez, J. M., Rajngewerc, M., Giudici, J. C., Furman, D. A., Luque, F., Alemany, L. A., and Martínez, M. V. (2021). pysentimiento: a python toolkit for opinion mining and social nlp tasks. *arXiv preprint arXiv:2106.09462*.

[Sandagiri et al. 2020a] Sandagiri, S., Kumara, B., and Kuhaneswaran, B. (2020a). Ann based crime detection and prediction using twitter posts and weather data. In *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, pages 1–5.

[Sandagiri et al. 2020b] Sandagiri, S., Kumara, B., and Kuhaneswaran, B. (2020b). Detecting crime related twitter posts using artificial neural networks based approach. In *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 5–10.

[Shoeibi et al. 2021] Shoeibi, N., Shoeibi, N., Hernández, G., Chamoso, P., and Corchado, J. M. (2021). Ai-crime hunter: An ai mixture of experts for crime discovery on twitter. *Electronics*, 10(24).

[Siddiqui et al. 2023] Siddiqui, T., Hina, S., Asif, R., Ahmed, S., and Ahmed, M. (2023). An ensemble approach for the identification and classification of crime tweets in the english language. *Computer Science and Information Technologies*, 4:149–159.

[Tedeschi et al. 2021] Tedeschi, S., Maiorca, V., Campolungo, N., Cecconi, F., and Navigli, R. (2021). Wikineural: Combined neural and knowledge-based silver data creation for multilingual ner. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 2521–2533.

[Vedova 2018] Vedova, D. (2018). O que é segurança publica. https://www.jusbrasil.com.br/artigos/o-que-e-seguranca-publica/586735267.

[Vieweg et al. 2010] Vieweg, S., Hughes, A. L., Starbird, K., and Palen, L. (2010). Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, page 1079–1088, New York, NY, USA. Association for Computing Machinery.

[Vo et al. 2020] Vo, T., Sharma, R., Kumar, R., Son, L. H., Pham, B. T., Tien Bui, D., Priyadarshini, I., Sarkar, M., and Le, T. (2020). Crime rate detection using social media of different crime locations and twitter part-of-speech tagger with brown clustering. *J. Intell. Fuzzy Syst.*, 38(4):4287–4299.

[WPR 2025] WPR (2025). Twitter/X Users by Country 2025. Online. World Population Review. https://worldpopulationreview.com/country-rankings/twitter-users-by-country.

[Zheng et al. 2018] Zheng, X., Han, J., and Sun, A. (2018). A survey of location prediction on twitter. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1652–1671.