Classificação de Dados Textuais Não Estruturados: Um Estudo de Caso na Área da Segurança Pública

Brenda Cardoso[1], Fantiny Santos[2], Angela Amador[1], Marisa de Andrade[1], Renato Torres[2], Nelson Neto[1]

[1]Programa de Pós-Graduação em Ciência da Computação (PPGCC-UFPA)
Rua Augusto Corrêa, 01 – Bairro Guamá – CEP 66075-110 – Belém – PA – Brasil

[2]Instituto de Ciências Exatas e Naturais – Universidade Federal do Pará (UFPA)
Rua Augusto Corrêa, 01 – Bairro Guamá – CEP 66075-110 – Belém – PA – Brasil

{brendakalline15,fantiny.santos,mm.marisamoreno}@gmail.com,
{angelaamador,renatohidaka,nelsonneto}@ufpa.br

Abstract. *The processing and classification of unstructured data are challenges in the information age. In the public security area, the lack of textual structuring of narratives in police reports (BOs) makes the precise categorization of crimes and the identification of the target audience even more complex. Thus, this paper proposes a method to speed up context classification in BOs through machine learning. The starting goal is to categorize crimes of insult directed or not at the LGBTQIA+ community based on reports from Pará State Police. The results highlight the potential applicability of the proposed approach in real and contextualized scenarios, contributing to the work of police authorities.*

Resumo. *O tratamento e a classificação de dados não estruturados são desafios na era da informação. Na segurança pública, a falta de estruturação textual das narrativas presentes nos boletins de ocorrência policial (BOs) torna ainda mais complexa a categorização precisa dos crimes e a identificação do público-alvo. Assim, este artigo propõe um método para agilizar a classificação de contexto em BOs por meio do aprendizado de máquina. A meta inicial é categorizar crimes de injúria direcionados ou não à comunidade LGBTQIA+ com base em relatos oriundos da Polícia Civil do Estado do Pará. Os resultados obtidos destacam a potencial aplicabilidade da abordagem proposta em cenários reais e contextualizados, contribuindo para o trabalho das autoridades policiais.*

1. Introdução

A quantidade de dados gerados cresce a um ritmo exponencial na chamada era da informação [George and Birla 2018]. Essa explosão de informações é vista em diversas áreas, como saúde, finanças e segurança pública. Porém, grande parte desses dados se encontra sem uma organização formal, como notícias e postagens em mídias sociais. Essa falta de estrutura limita a análise e a extração de revelações valiosas, exigindo dos usuários um trabalho tedioso de leitura e interpretação [Mallek et al. 2020].

Lidar com dados não estruturados é uma necessidade e, ao mesmo tempo, uma dificuldade para os órgãos de segurança pública. Boletins de ocorrência policial (BOs) e imagens de câmeras de segurança são exemplos de dados que podem conter informações cruciais para prevenir delitos e investigar crimes. No entanto, a exploração eficiente de

dados não estruturados é um desafio [Mallek et al. 2020] e exige ferramentas de análise avançadas e capazes de lidar com a natureza complexa e desorganizada desses dados.

A identificação de contexto é um típico problema que precisa da análise de dados não estruturados. Por exemplo, a ausência de estruturação nos relatos textuais dos BOs dificulta a precisão no que tange a classificação dos crimes, a identificação do público-alvo e a compreensão das motivações por trás dos delitos. Essa carência de informação limita a capacidade das autoridades de prevenir e investigar os crimes com eficiência. O processamento de linguagem natural (PLN) [Pinheiro et al. 2010] e o aprendizado de máquina (ML) [Sakhare and Joshi 2014] surgem como ferramentas valiosas ao ajudarem na estruturação e categorização dos dados, além da extração de conhecimento relevante.

Assim, de um lado, tem-se à disposição modernas técnicas de pré-processamento e classificação de dados textuais, do outro, grandes volumes de dados não estruturados produzidos diariamente por órgãos governamentais, como a Polícia Civil (PC), cujo trabalho diário resulta em um gigantesco número de BOs. Por exemplo, os crimes de injúria, que é o contexto deste trabalho, não possuem distinção de público-alvo (p.e. contra mulheres, pessoas da comunidade LGBTQIA+, entre outros) no sistema da PC do Pará.

A análise de dados criminais é uma fonte essencial de informação para auxiliar os tomadores de decisões sociais e políticas no que diz respeito à alocação de recursos de segurança pública [Rodrigues et al. 2023]. No entanto, a falta de estruturação e a categorização precária desses dados textuais prejudicam a qualidade da análise, seja ela feita de maneira automática ou manual. O desperdício de tempo associado também é um fator que precisa ser levado em consideração.

Dado o cenário atual de dificuldade de tratamento de dados textuais não estruturados, este artigo testa a seguinte hipótese: *O desenvolvimento de modelos de aprendizado de máquina é capaz de apresentar resultados satisfatórios para a classificação de boletins de ocorrência policial, levando em consideração o contexto do relato*. Dessa forma, o objetivo geral deste trabalho é agilizar o processo de classificação de contexto em BOs usando algoritmos de ML. Como esforço inicial, a ideia é construir modelos computacionais supervisionados para classificação binária de crimes de injúria voltados ou não à comunidade LGBTQIA+.

As seções subsequentes apresentam os trabalhos relacionados, a metodologia adotada, bem como, os resultados e conclusões da pesquisa.

2. Trabalhos Relacionados

A busca automática por trabalhos correlatos foi realizada em três renomados repositórios: *IEEE Xplore*, *ACM Digital Library* e *BioMed Central*. Os critérios de seleção incluíram artigos escritos na língua inglesa, publicados entre 2017 e 2023 e que apresentam algum modelo computacional para classificação de texto não estruturado de documentos relacionados à segurança pública. Além disso, buscas manuais foram feitas em conferências e periódicos da Sociedade Brasileira de Computação no intuito de identificar pesquisas que abordam o contexto brasileiro.

Em [Kuang et al. 2017], ou autores propõem um método para classificar crimes usando descrição textual de eventos e modelagem de tópicos. A técnica de aprendizado de máquina empregada foi a fatorização de matriz não negativa (NMF), que é uma abor-

dagem não supervisionada. O método foi aplicado em dados do Departamento de Polícia de Los Angeles, EUA, de 2009 a 2014. O modelo identificou distinções textuais entre crimes violentos e de propriedade, mas não foi considerado suficiente para classificação automática. A pesquisa destaca o potencial da modelagem de tópicos para auxiliar na tomada de decisões para a redução de crimes.

Outra abordagem não supervisionada, dessa vez a *Latent Dirichlet Allocation* (LDA), foi usada em [Birks et al. 2020]. Com base em transcrições livres de gravações de áudio feitas pela polícia, os autores apresentam um método para analisar roubos residenciais no Reino Unido, que utiliza modelagem de tópicos e aprendizado de máquina para, respectivamente, agrupar automaticamente os dados e identificar crimes de acordo com categorias existentes. Os autores acreditam que esse método pode automatizar tarefas que demandam muito tempo de analistas criminais.

Ferramentas de mineração de texto foram exploradas por [Palad et al. 2019] para extrair informações e encontrar padrões em um conjunto de dados de golpes online nas Filipinas. Os algoritmos supervisionados J48, *Naive Bayes* e *Sequential Minimal Optimization* (SMO) foram usados para construir modelos de classificação e seus resultados foram comparados. O J48 obteve a melhor precisão e menor taxa de erro, seguido por *Naive Bayes* e SMO, nessa ordem. Os resultados foram validados por investigadores de crimes cibernéticos. A pesquisa demonstra como a análise preditiva é capaz de auxiliar na análise e identificação de incidentes de fraudes online.

O estudo de [Nasr et al. 2022] propõe um sistema para auxiliar a classificação automática de textos em árabe provenientes de linhas diretas de emergência da polícia. O objetivo é reduzir o risco de erros e decisões equivocadas na categorização manual feita pelos operadores. A pesquisa considerou 13 tipos de crimes e avaliou diversas combinações de técnicas de PLN e ML. No geral, o modelo AraBERT obteve o melhor desempenho, com 92,3% de precisão. Contudo, quando combinado com a técnica de vetorização TF-IDF, os modelos supervisionados *Support Vector Machines* (SVM) e *Random Forest* obtiveram os melhores resultados, atingindo 90,1% de precisão.

Em [Gusmão et al. 2021], técnicas de PLN e ML foram aplicadas em denúncias criminais, oriundas do aplicativo Disque Denúncia RJ. O artigo apresenta um método para classificação automática do tipo de denúncia (tráfego de drogas, homicídios, crimes contra mulheres, entre outros), visando otimizar o tempo de análise do conteúdo das mensagens, que estão escritas em linguagem informal, o que foi um grande desafio para os autores. Os resultados obtidos com o algoritmo SVM foram positivos e demostram que o classificador tem boa capacidade de generalização, atingindo uma precisão de 76,11%.

A maioria dos trabalhos correlatos categorizam crimes de natureza geral, como os violentos e contra o patrimônio. Este trabalho se destaca por combinar técnicas de PLN e ML utilizando uma abordagem supervisionada para lidar com a complexidade dos dados textuais nos BOs, especialmente ao classificar um tipo específico de crime: injúria contra pessoas LGBTQIA+. Outra distinção importante é a utilização de dados oriundos da PC do Estado do Pará. Essas especificidades assumem caráter prático e possibilitam uma análise mais realista e contextualizada do problema em questão.

3. Materiais e Métodos

Nesta seção, serão apresentadas as tarefas realizadas no desenvolvimento do classificador binário. Conforme o fluxograma apresentado na Figura 1, quatro grandes etapas foram realizadas, incluindo a coleta e entendimento dos dados, o pré-processamento dos dados, a modelagem e, finalmente, a fase de avaliação.

Essas etapas foram adaptadas da metodologia de análise de dados conhecida como *Cross-Industry Standard Process for Data Mining* (CRISP-DM). Segundo [Hotz 2023], esse modelo serve de base para um processo de ciência de dados, contando com seis fases: (i) entendimento do negócio; (ii) compreensão de dados; (iii) preparação de dados; (iv) modelagem; (v) avaliação; e (vi) implantação.

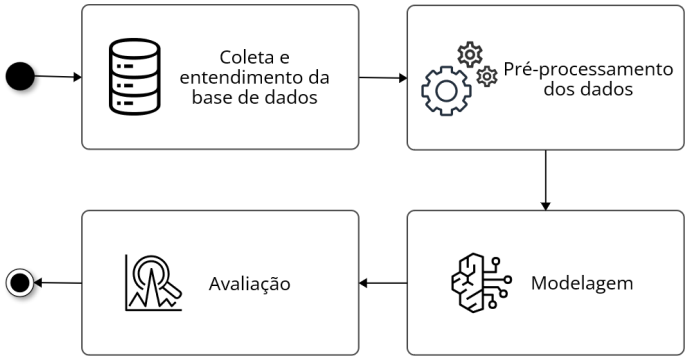

Figura 1. A metodologia utilizada é uma adaptação do modelo de referência *Cross-Industry Standard Process for Data Mining* [Hotz 2023].

3.1. Construção da Base de Dados

Este trabalho analisou relatos de vítimas de injúria registrados em boletins de ocorrência da Polícia Civil do Pará (PC/PA), entre janeiro de 2022 e setembro de 2023, conforme descrito na Tabela 1. Os dados foram obtidos junto à Diretoria Estadual de Combate ao Crime Cibernético (DECCC), que autorizou esta pesquisa documental mediante a assinatura de um termo de confiabilidade e sigilo.

Tabela 1. Resumo dos registros de crimes disponibilizados pela DECCC.

Tipo	Quantidade
Violência Contra a Mulher	1.022
Injúria	8.434
Roubo	1.022
Crime Virtual	1.022

A base de dados para a classificação de crimes contra pessoas LGBTQIA+ foi disponibilizada pela PC/PA juntamente com os crimes de injúria, já que os crimes de caráter homofóbico não possuem uma rotulação específica que os identifique. Os dados incluem 8.434 registros de injúria, cada um com 20 atributos. Os atributos incluem informações como local, data, hora, natureza, motivação e relato do crime. Importante salientar que a base de dados foi anonimizada pela própria DECCC, impossibilitando a identificação de dados pessoais e, por conseguinte, preservando a identidade dos envolvidos.

Esta pesquisa baseia-se em aprendizado supervisionado, logo, a rotulação dos dados é necessária no processo de treinamento do modelo computacional. Durante o processo de rotulação, surgiram dúvidas sobre quais são os critérios que caracterizam um crime como homofóbico. Por isso, primeiramente, especialistas na área do Direito foram consultados e suas instruções levadas em consideração. Em seguida, adotou-se a metodologia binária para realizar a rotulação dos dados, ou seja, os relatos (atributo) foram rotulados como sendo ou não casos de violência contra pessoas LGBTQIA+. Quatro julgadores, pesquisadores acadêmicos, realizaram a tarefa em duas duplas. No caso de divergência entre os dois julgadores, um terceiro era consultado.

Na Figura 2, é possível visualizar o fluxo seguido para rotular os dados. Cada dupla recebeu metade dos registros de injúria concedidos (4.217), porém, apenas 2.000 relatos foram analisados e ao todo 990 manualmente rotulados, sendo 349 deles contra a comunidade LGBTQIA+. Cabe ressaltar que a taxa de concordância entre os julgadores foi de 90% e que apenas sete atributos foram mantidos, já que os considerados irrelevantes para a pesquisa foram descartados, como latitude, longitude, e outros.

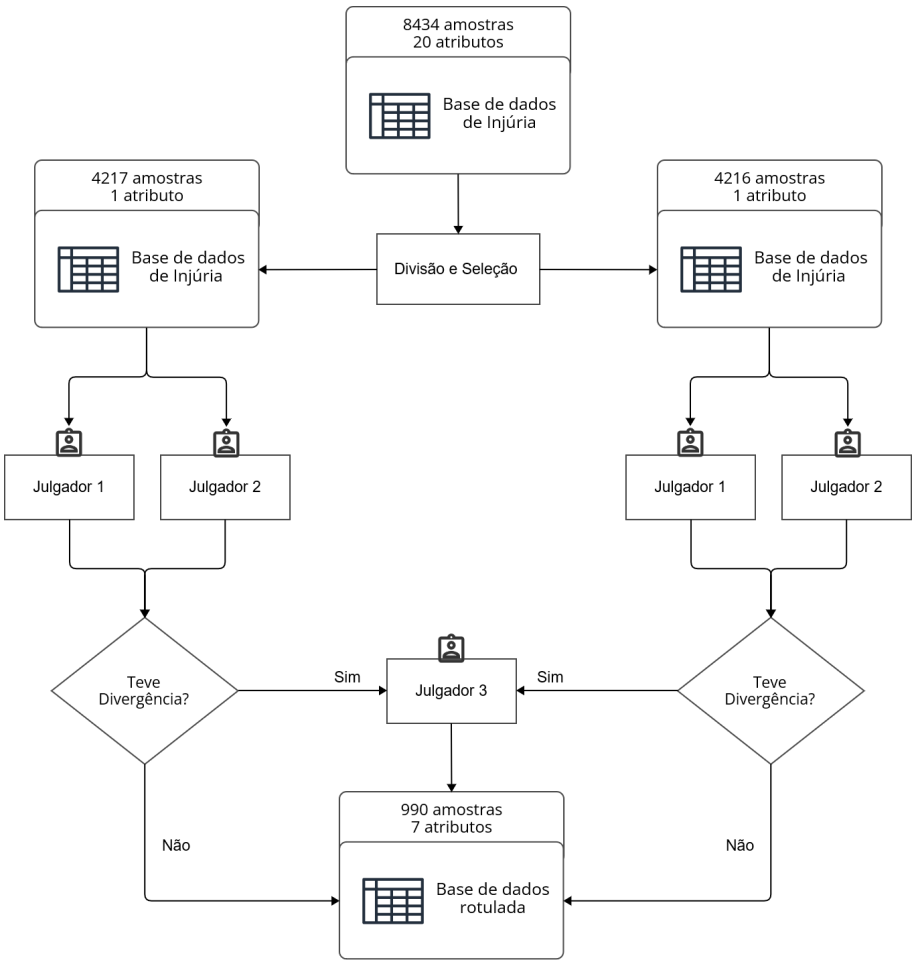

Figura 2. Fluxograma do processo de rotulação dos dados relacionados ao crime de injúria obtidos junto a Polícia Civil do Estado do Pará.

Mais tarde, uma segunda remessa contendo 184 registros com 18 atributos cada foi disponibilizada pela DECCC, mas dessa vez com dados rotulados internamente por poli-

ciais civis. Os 51 registros rotulados como crime de injúria contra pessoas LGBTQIA+ foram, então, adicionados aos 349 anteriormente rotulados, totalizando 400 amostras. Para finalizar a construção da base de dados, foram adicionadas 400 amostras de outros contextos. Essas amostras foram escolhidas aleatoriamente do conjunto rotulado neste trabalho, rejeitando aquelas classificadas como violência contra a mulher.

3.2. Pré-Processamento dos Dados

As narrativas baseadas em textos são consideravelmente ruidosas, no sentido de conterem erros de digitação, abreviações, *tags*, entre outros caracteres que dificultam o entendimento do conteúdo escrito. Assim, algumas tarefas de pré-processamento de texto foram executadas nos relatos visando obter resultados mais confiáveis e menos sensíveis a esses tipos de ruídos. A estratégia adotada neste trabalho pode ser visualizada na Figura 3.

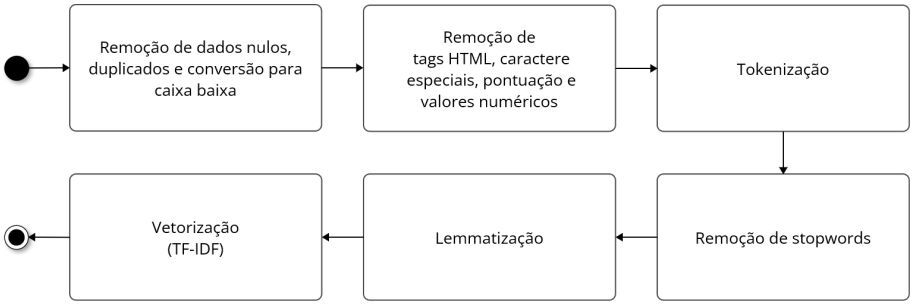

Figura 3. Fluxograma das tarefas realizadas no pré-processamento dos dados.

Após a limpeza dos dados (remoção dos registros nulos ou duplicados), iniciou-se a fase de formatação, no intuito de simplificar o texto e reduzir o número de palavras únicas. A seguir, a lista das tarefas realizadas: **Lowercasing**: Converte todos os caracteres do texto em minúsculos. Para a conversão em caixa baixa, usou-se o método *srt.lower()* da biblioteca *pandas*[1] do Python. **Remoção de pontuação e caracteres especiais**: Envolve a remoção de todos os sinais de pontuação do texto, incluindo símbolos como vírgulas, pontos, dois-pontos, ponto-e-vírgula e outros. Abrange também a remoção de caracteres não alfanuméricos do texto, incluindo símbolos como *hashtags*, arrobas, cifrões e outros caracteres especiais que não sejam letras ou números. **Normalização numérica**: Envolve a conversão dos símbolos numéricos para um formato padrão. Para isso, usou-se o método *sub()* da biblioteca *RegEx*[2] do Python para converter [0-9+] em vazio. Os algarismos foram removidos porque entendeu-se que eles não adicionam ganho de informação para o tipo de classificação almejada. **Tokenização**: Consiste na segmentação do texto em unidades básicas, ou *tokens*. Neste trabalho, um *token* equivale a uma palavra, tendo espaços em branco como delimitadores. Para isso, usou-se o método *word_tokenize* da biblioteca NLTK[3] do Python, que aciona o *tokenizer TreebankWordTokenize* junto com o *PunktSentenceTokenizer*. **Remoção de *stopwords***: Envolve a retirada de palavras comuns do texto, como artigos e preposições. Para garantir uma maior efetividade, usou-se a lista de *stopwords* do Português Brasileiro das bibliotecas NLTK e spaCy[4] combina-

[1]https://pandas.pydata.org/
[2]https://docs.python.org/pt-br/3/library/re.html
[3]https://www.nltk.org/
[4]https://spacy.io/

das. **Lematização**: Compreende a redução de palavras à sua forma canônica, conhecida como lema. Para realizar a lematização do texto, usou-se o modelo pré-treinado *pt_core_news_sm* e o método *.lemma_* da biblioteca spaCy do Python.

Por fim, a representação numérica dos dados textuais (ou vetorização) foi realizada por meio da abordagem TF-IDF, que monta a entrada para o algoritmo de aprendizado de máquina, transformando *tokens* (palavras isoladas) em informações numéricas [Ramos et al. 2003]. Neste trabalho, os *tokens* presentes em menos de 1% e os que aparecem em mais de 90% do *corpus* foram ignorados.

3.3. Modelagem

Quando o conjunto de dados é pequeno, a divisão habitual de treino e teste pode não ser útil. Mesmo que as partes sejam criadas de forma aleatória, pode ocorrer uma seleção de valores mais característicos dependendo da origem da casualidade. A escassez de dados rotulados pode resultar em agrupamentos de teste insuficientes para certificar a capacidade de generalização do modelo [Shalev-Shwartz and Ben-David 2014].

Uma solução para esse problema é realizar várias divisões de treino e teste no conjunto de dados. Assim, o modelo é treinado e avaliado com diferentes subconjuntos de dados. Uma possibilidade para se fazer isso é por meio da validação cruzada, onde o conjunto de dados é dividido em várias partes conhecidas como partições (ou *folds*). Neste estudo, a validação cruzada foi realizada com 10 partições, onde nove são usadas para treinamento e uma para validação. Esse processo é, então, repetido 10 vezes, com cada partição sendo usada uma vez para validação.

Os seguintes algoritmos de ML foram escolhidos para a construção dos modelos computacionais: *Support Vector Machine*, *Random Forest*, *Logistic Regression* e *Gradient Boosting*. Os algoritmos foram importados da biblioteca *Sklearn*[5], versão 1.3.2, com seus parâmetros padrões. A opção por esses algoritmos se deu, principalmente, pelo uso recorrente deles em trabalhos correlatos, tanto na classificação de textos [Albrecht et al. 2020], como em outras atividades correlatas [Wei 2023]. Os parâmetros padrões da biblioteca foram usados por ser a solução mais rápida e porque o objetivo deste trabalho não é definir o melhor algoritmo, mas, sim, avaliar se o aprendizado de máquina consegue de alguma forma agilizar o processo de classificação de BOs dado o contexto exposto.

4. Avaliação

Este seção descreve dois experimentos em classificação de BOs usando os materiais e métodos adotados. Em cada experimento, serão apresentados os resultados da aplicação de algoritmos aprendizado de máquina supervisionado no problema estudado.

4.1. Experimento 1: Classificação de relatos de crimes contra a comunidade LGBTQIA+

Este experimento teve o intuito de treinar um modelo computacional para estimar quando um relato de injúria é do contexto de crime contra a comunidade LGBTQIA+. Para isso, a base de dados com 800 registros descrita na Seção 3.1 foi utilizada. As amostras foram divididas igualmente entre o contexto LGBTQIA+ (rótulo 1) e relatos fora desse contexto

[5]https://scikit-learn.org/stable/index.html

(rótulo 0). Em seguida, a base de dados foi separada em conjuntos de treino e teste, cujos percentuais foram definidos em 70% e 30%, respectivamente. A quantidade de registros de cada conjunto de dados é mostrada na Tabela 2.

Tabela 2. Divisão da base de dados em treino e teste.

	Rótulo 0 - Outros Contextos	Rótulo 1 - LGBTQIA+	Total
Treino	285	275	560
Teste	115	125	240
Total	400	400	800

Então, modelos computacionais foram treinados com os quatro algoritmos escolhidos e a matriz de confusão gerada para cada um deles pode ser vista na Figura 4. É possível verificar que, no modelo *Random Forest*, por exemplo, 110 dos 240 relatos da base de teste foram classificados corretamente como sendo do contexto LGBTQIA+. Por sua vez, no modelo *Gradient Boosting*, 112 das 240 amostras de teste foram classificadas corretamente como não sendo do contexto LGBTQIA+. Já o modelo SVM apresentou 5 casos de falso positivo e 37 casos de falso negativo.

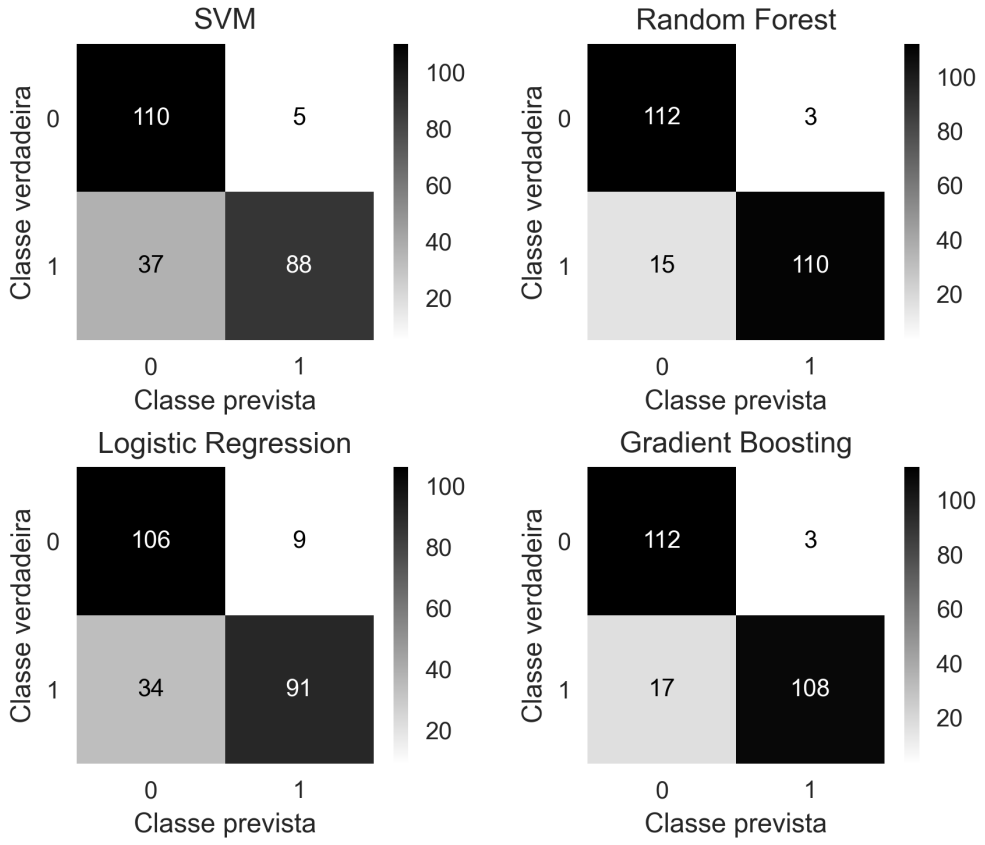

Figura 4. Matriz de confusão resultante do experimento 1.

Um ponto importante que deve ser observado, considerando o problema motivador deste trabalho, é que os modelos computacionais construídos foram especialistas exatamente onde esperava-se que eles fossem, ou seja, na minimização dos falsos positivos.

Desse modo, o(a) investigador(a) policial tem mais confiança para desempenhar seu trabalho, pois, sabe-se que a chance da solução classificar crimes no contexto desejado, mas que na verdade não são, é baixa.

A validação cruzada (10-*folds*) foi utilizada com o objetivo de avaliar o comportamento dos algoritmos quando submetidos a um processo de generalização. De acordo com os valores mostrados na Tabela 3, percebe-se que o modelo com a generalização mais estável foi o *Grandient Boosting*, com aproximadamente 88% de *F1-Score* médio.

Tabela 3. Validação cruzada do experimento 1.

	F1-Score	**Desvio Padrão**
Support Vector Machine	0.78754	0.06213
Random Forest	0.86648	0.03452
Logistic Regression	0.79765	0.06607
Gradient Boosting	**0.88251**	**0.02385**

No geral, os resultados alcançados apontam que os classificadores binários construídos, mesmo sem ajuste de parâmetros, podem ser ferramentas úteis para a detecção de crimes de injúria contra a comunidade LGBTQIA+ considerando um cenário local.

4.2. Experimento 2: Classificação de relatos de crimes contra a comunidade LGBTQIA+ com ruído nos dados

O objetivo do experimento 2 foi avaliar o impacto da inclusão de relatos rotulados como violência doméstica e familiar contra a mulher (VDFCM) no desempenho dos modelos treinados para identificar crimes contra a comunidade LGBTQIA+. A similaridade entre esses dois contextos levanta a seguinte questão: A adição de relatos de violência contra a mulher atrapalha a identificação de crimes homofóbicos?

Este ensaio utilizou os mesmos 400 relatos de crimes de injúria contra a comunidade LGBTQIA+ (rótulo 1) usados no experimento 1. A diferença entre os experimentos está nos outros 400 registros (rótulo 0). Agora, apenas 264 relatos pertencem a outros contextos e 136 são específicos de VDFCM. Esses últimos possuem contexto similar aos relatos homofóbicos e não foram vistos no experimento 1. Em seguida, os conjuntos de treino (70%) e teste (30%) foram montados como pode ser visto na Tabela 4.

Tabela 4. Divisão da base de dados em treino e teste.

	Rótulo 0 - Incluindo VDFCM	**Rótulo 1 - LGBTQIA+**	**Total**
Treino	285	275	560
Teste	115	125	240
Total	400	400	800

Então, modelos computacionais foram treinados com os mesmos algoritmos usados no experimento 1 e a matriz de confusão resultante de cada um deles é apresentada na Figura 5. Devido à adição de relatos de contexto similar, como esperado, os casos de falso positivo cresceram comparados com o experimento 1, mas, ainda assim, foi a situação que apresentou os menores valores na matriz de confusão, o que é um ponto positivo conforme explicado anteriormente.

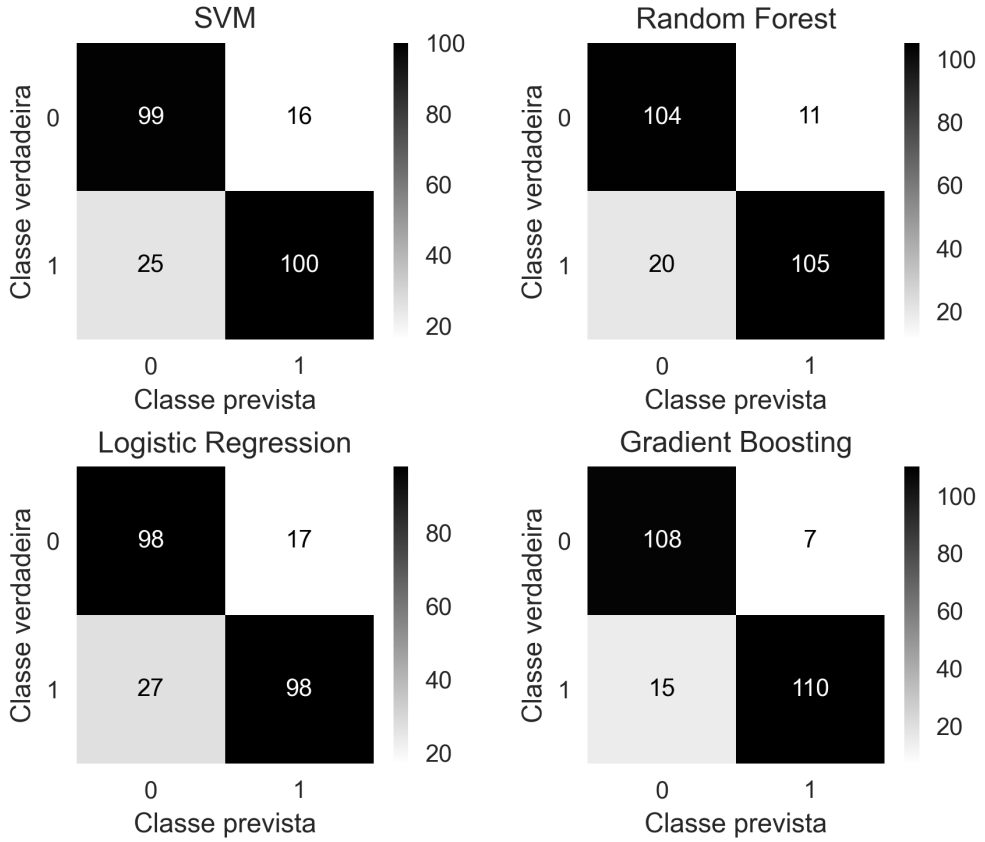

Figura 5. Matriz de confusão resultante do experimento 2.

Os resultados da validação cruzada mostraram novamente o algoritmo *Grandient Boosting* com a generalização mais estável, com cerca de 90% de *F1-Score* médio, conforme a Tabela 5. Buscando mais precisão, o teste de significância estatística entre os dois experimentos foi feito a partir da observação da métrica *F1-score* coletada em três rodadas de validação cruzada, resultando, assim, em 30 valores de *F1-score* para cada modelo computacional. A densidade dos modelos foi calculada a partir desses valores e as funções de densidade não foram perfeitamente caracterizadas como uma distribuição Gaussiana (vide Figura 6). Por isso, usou-se o teste não paramétrico de Kolmogorov-Smirnov.

Tabela 5. Validação cruzada do experimento 2.

	F1-Score	**Desvio Padrão**
Support Vector Machine	0.80029	0.06696
Random Forest	0.86645	0.05091
Logistic Regression	0.79303	0.07350
Gradient Boosting	**0.90241**	**0.03978**

Em função dos contextos LGBTQIA+ e VDFCM compartilharem a temática de violência de gênero, motivada por preconceito e discriminação, a inclusão de relatos de VDFCM proporcionou uma diferença significativa nos modelos *Random Forest* e *Gradient Boosting* com p-value = 0.00000127500603424 e p-value = 0.00000025000118318, respectivamente.

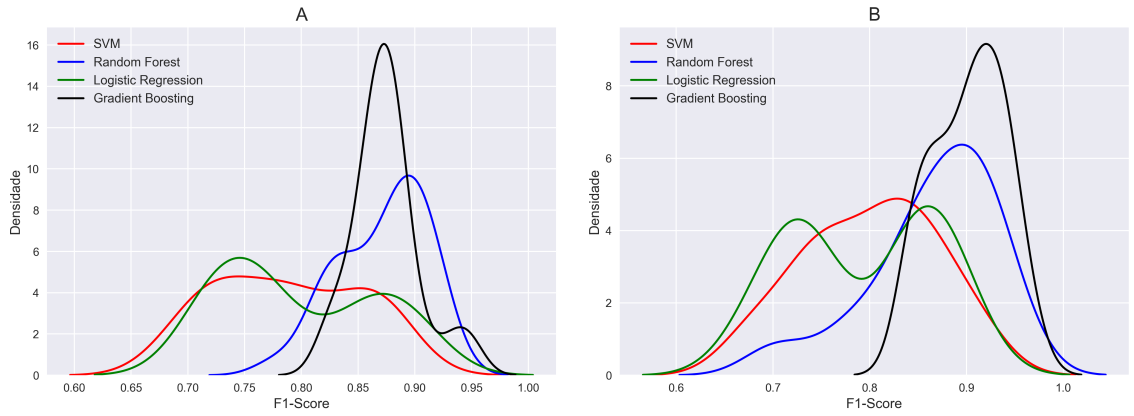

Figura 6. Estimativa de densidade dos modelos computacionais. Gráficos A e B referentes aos experimentos 1 e 2, respectivamente.

5. Conclusões

Esta pesquisa tem como propósito investigar o desempenho de modelos de classificação automática de crimes de injúria direcionados à comunidade LGBTQIA+ a partir de dados reais. Os resultados alcançados nesse primeiro esforço são promissores ao indicarem que o aprendizado de máquina pode agilizar o processo de categorização de boletins de ocorrência policial considerando o contexto abordado.

Em função dos contextos LGBTQIA+ e VDFCM compartilharem a temática de crimes motivados por preconceito e discriminação, os resultados obtidos revelam discrepâncias no desempenho dos modelos de classificação binária quando esses contextos são misturados. No primeiro experimento, que abordou exclusivamente crimes relacionados à comunidade LGBTQIA+, o algoritmo *Gradient Boosting* alcançou a generalização mais estável ao observarmos os resultados da validação cruzada.

O segundo ensaio, que envolveu a combinação de crimes LGBTQIA+ e VDFCM, apontou um sensível aumento nos casos de falso positivo, o que revela a necessidade de investigações futuras voltadas para o desenvolvimento de modelos específicos para cada contexto, levando em conta suas características e nuances. Ressalta-se que o algoritmo *Gradient Boosting* apresentou novamente a generalização mais estável e que a diferença da métrica *F1-score* entre os dois experimentos foi comprovada estatisticamente. Em todo caso, mais testes precisam ser feitos, considerando o aumento da base de dados, a validação dos dados rotulados por especialistas da área e o uso de algoritmos de ML com diferentes tipos de abordagens.

A análise detalhada da dinâmica da violência, com foco em motivações de ódio, pode ser um caminho promissor para pesquisas futuras. O emprego de ferramentas de explicabilidade, que possibilitem uma exploração mais profunda dos dados, pode ajudar na identificação de padrões que levam a esse tipo de violência, ou mesmo que ela encontra-se em andamento, averiguando de forma independente nichos da sociedade, principalmente os grupos vulneráveis, como as mulheres e a comunidade LGBTQIA+.

Referências

Albrecht, J., Ramachandran, S., and Winkler, C. (2020). *Blueprints for Text Analytics Using Python*.

Birks, D., Coleman, A., and Jackson, D. (2020). Unsupervised identification of crime problems from police free-text data. *Crime Science*, 9(1):18.

George, L. E. and Birla, L. (2018). A study of topic modeling methods. In *2018 second international conference on intelligent computing and control systems (iciccs)*, pages 109–113. IEEE.

Gusmão, C., Figueiredo, K., and Brito, W. A. (2021). Técnicas de processamento de linguagem natural em denúncias criminais: Automatizaçao e classificaçao de texto em português coloquial. In *Anais do XLVIII Seminário Integrado de Software e Hardware*, pages 172–182. SBC.

Hotz, N. (2023). What is CRISP DM? - Data Science Process Alliance — datascience-pm.com. `https://www.datascience-pm.com/crisp-dm-2/`. [Accessed 31-01-2024].

Kuang, D., Brantingham, P. J., and Bertozzi, A. L. (2017). Crime topic modeling. *Crime Science*, 6(1):1–20.

Mallek, M., Fournier, S., Guetari, R., Espinasse, B., and Chaari, W. L. (2020). An unsupervised approach for precise context identification from unstructured text documents. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 821–826. IEEE.

Nasr, B., Chamoun, M., and Steyaert, J. M. (2022). Optimizing the process of police hotlines. In *2022 IEEE 1st Industrial Electronics Society Annual On-Line Conference (ONCON)*, pages 1–6. IEEE.

Palad, E. B. B., Tangkeko, M. S., Magpantay, L. A. K., and Sipin, G. L. (2019). Document classification of filipino online scam incident text using data mining techniques. In *2019 19th International Symposium on Communications and Information Technologies (ISCIT)*, pages 232–237. IEEE.

Pinheiro, V., Furtado, V., Pequeno, T., and Nogueira, D. (2010). Natural language processing based on semantic inferentialism for extracting crime information from text. In *2010 IEEE International Conference on Intelligence and Security Informatics*, pages 19–24.

Ramos, J. et al. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

Rodrigues, A., González, J. A., and Mateu, J. (2023). A conditional machine learning classification approach for spatio-temporal risk assessment of crime data. *Stochastic Environmental Research and Risk Assessment*, pages 1–14.

Sakhare, N. N. and Joshi, S. A. (2014). Classification of criminal data using j48-decision tree algorithm. *IFRSA International Journal of Data Warehousing & Mining*, 4.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Wei, L. (2023). Genetic algorithm optimization of concrete frame structure based on improved random forest. In *2023 International Conference on Electronics and Devices, Computational Science (ICEDCS)*, pages 249–253.