Avaliação Preliminar de Técnicas de PLN para Classificação de Relatos em Boletins de Ocorrência Policial

Ryan F. de Sousa[1], **Raimundo S. Moura**[1]

[1]Departamento de Computação – Universidade Federal do Piauí (UFPI)
64.049-550 – Teresina – PI – Brazil

{ryanfsousa,rsm}@ufpi.edu.br

Abstract. *This work investigates the application of Natural Language Processing (NLP) techniques for analyzing textual narratives from Police Reports (Boletins de Ocorrência - BOs) provided by the Public Security Secretariat of Piauí State, Brazil. Manual analysis of these reports is complex and time-consuming. This initial phase focuses on automatically classifying the crime type based on the unstructured textual description, a fundamental step for future consistency checks and information extraction. We evaluate three approaches: Random Forest with TF-IDF features, a hybrid CNN-LSTM network with pre-trained GloVe embeddings, and a fine-tuned Large Language Model (LLM), Qwen 2.5 7B, using QLoRA. Preliminary results on a balanced dataset show the fine-tuned LLM achieved the best performance (F1-Score: 0.69), surpassing Random Forest (0.61) and CNN-LSTM (0.59). Data quality issues were identified as a potential bottleneck, suggesting future work should focus on data cleaning and refining extraction methods using LLMs.*

Resumo. *Este trabalho investiga a aplicação de técnicas de Processamento de Linguagem Natural (PLN) para análise de relatos textuais de Boletins de Ocorrência (BOs) da Secretaria de Segurança Pública do Estado do Piauí. A análise manual destes relatos é complexa e morosa. Esta fase inicial foca na classificação automática da natureza da ocorrência a partir da descrição textual não estruturada, um passo fundamental para futuras verificações de consistência e extração de informações. Avaliamos três abordagens: Random Forest com TF-IDF, uma rede híbrida CNN-LSTM com embeddings GloVe pré-treinados, e um Grande Modelo de Linguagem (LLM) Qwen 2.5 7B ajustado via QLoRA. Resultados preliminares em um dataset balanceado indicam que o LLM ajustado obteve o melhor desempenho (F1-Score: 0.69), superando o Random Forest (0.61) e a CNN-LSTM (0.59). Questões de qualidade dos dados foram identificadas como um possível gargalo, sugerindo que trabalhos futuros devem focar na limpeza dos dados e no refinamento de métodos de extração com LLMs.*

1. Introdução

A análise de grandes volumes de dados textuais gerados em Boletins de Ocorrência (BOs) representa um desafio significativo para os órgãos de segurança pública. A extração eficiente de informações e a verificação da consistência dos relatos são cruciais para a gestão e tomada de decisão. No Estado do Piauí, os registros online via Delegacia Virtual (2025)

(DEVIR) geram um volume expressivo de dados, incluindo descrições detalhadas dos fatos em formato de texto não estruturado. O processo manual de análise desses textos é intensivo em mão de obra, lento e sujeito a inconsistências.

Diante desse cenário, este trabalho insere-se no contexto de investigação da aplicação de técnicas de Processamento de Linguagem Natural (PLN) para automatizar e auxiliar na análise desses relatos. O objetivo geral é desenvolver e avaliar métodos para extração de informações relevantes e verificação de coerência nos BOs.

Inicialmente, o foco foi direcionado a uma tarefa fundamental: a classificação automática da natureza da ocorrência com base exclusivamente no relato textual fornecido pelo comunicante. A correta classificação do tipo de crime descrito no texto livre é um pré-requisito essencial para etapas subsequentes, como a verificação de inconsistências entre o relato e a natureza selecionada no formulário, e a extração direcionada de informações específicas, como objetos subtraídos, dados de suspeitos, modus operandi e outras.

Este artigo apresenta a metodologia empregada, os modelos de aprendizado de máquina e PLN avaliados, incluindo abordagens clássicas e Grandes Modelos de Linguagem (do inglês: *Large Language Models* - LLMs), e os resultados preliminares obtidos na tarefa de classificação, utilizando um conjunto de dados reais fornecidos pela Secretaria de Segurança Pública do Estado do Piauí.

2. Trabalhos Relacionados

Na literatura especializada, muitos trabalhos têm explorado o uso de IA na área de segurança pública, incluindo artigos e monografias de Trabalho de Conclusão de Curso (TCC) e Dissertação de Mestrado.

Muitos pesquisadores brasileiros também abordaram o tema. Amorim and Pereira (2019) usaram algoritmos clássicos (C4.5, CART, k-NN, SVM, Redes Neurais, Ripper, RF) para classificar automaticamente ocorrências policiais, incluindo o texto dos relatos. Eles enfrentaram desafios com o alto número de classes (381 originais) e reportaram acurácia de até 99% para C4.5 e Ripper em subconjuntos menores, mas com grande variabilidade e resultados baixos em outros cenários, destacando a dificuldade com muitas classes e poucas instâncias.

Castro (2020) focou na predição de padrões criminais (roubo/furto) usando fontes heterogêneas (oficiais e não oficiais) em Belo Horizonte, sem utilizar o texto das ocorrências. Explorou k-NN, SVM, RF, XGBoost e LSTM, com este último alcançando a maior acurácia (91%) na previsão de tendências e ocorrências por tipo e região, usando atributos como data, local e tipo de crime.

Anjos Junior and et al. (2020) aplicaram AM não supervisionado (DBSCAN) para mapear o comportamento de roubos e furtos de veículos em João Pessoa, utilizando apenas dados de data, hora e localização geográfica. O trabalho identificou hotspots[1] e bairros com maiores taxas de ocorrência, sem usar o conteúdo textual dos BOs.

Matos et al. (2022) desenvolveram um classificador supervisionado (CNN) para BOs do Pará (2019-2021), utilizando o texto. Com uma base grande e ruidosa (1.3M

[1]Hotspots são locais que concentram altas ocorrências de atividades criminais.

registros, 463 classes após pré-processamento), alcançaram acurácia geral de 78%, com mais de 85% para classes como roubo e furto, mas limitada pela quantidade de classes e desbalanceamento.

Kremer (2023) analisou roubos e furtos de veículos em Porto Alegre com AM supervisionado (k-NN, Árvore de Decisão, RF, GB) e não supervisionado (K-Means), sem usar o texto. K-Means ajudou a identificar padrões geográficos e de valor dos veículos. GB obteve a melhor performance na previsão de recuperação de veículos (64% de acurácia) e do local de recuperação (67%), usando atributos como tipo, local, data e características do veículo.

Esses trabalhos mostram diferentes abordagens, com ou sem o uso do texto narrativo, e desafios relacionados à qualidade dos dados, número de classes e desbalanceamento, problemas também pertinentes à presente pesquisa.

3. Metodologia

A metodologia adotada envolveu o pré-processamento dos dados, a definição do conjunto de dados para treinamento e teste, e a implementação e avaliação de três diferentes abordagens de classificação.

Utilizou-se um conjunto de dados sigilosos de BOs registrados na Delegacia Virtual (DEVIR) do Estado Piauí, abrangendo os anos de 2022 a 2024. Cada registro contém, entre outras informações, o relato textual da ocorrência e dois campos de classificação: "Natureza", preenchida pelo comunicante, e "Natureza Ajustada", revisada e confirmada por um homologador. A "Natureza Ajustada" foi utilizada como o rótulo (classe) verdadeiro para a tarefa de classificação.

Realizou-se um processo de agrupamento de naturezas semanticamente similares, como diferentes formas de registrar "Perda ou extravio", para consolidar as classes. Dentre as naturezas resultantes, foram selecionadas para o estudo apenas as que possuíam, no mínimo, 500 amostras após o agrupamento. A Tabela 1 detalha as contagens originais destas categorias antes do agrupamento e seleção final para o balanceamento.

Devido ao desbalanceamento natural entre as classes selecionadas (após agrupamento e filtro de contagem mínima, conforme descrito no texto original), aplicou-se a técnica de *undersampling* aleatório sobre as majoritárias para criar um conjunto de dados de treinamento e teste balanceado, consistindo em 514 amostras para treinamento e 25 ou 26 amostras para teste por classe.

Considerando a natureza sensível dos dados, todos os experimentos foram conduzidos em ambiente local, utilizando uma máquina com processador i5-12400F, 32GB de RAM e uma GPU NVIDIA RTX 3060 com 12GB de VRAM. As principais ferramentas e bibliotecas utilizadas foram *Python* 3, *NLTK* (Bird et al., 2009) (para pré-processamento básico como remoção de *stop words* e pontuação), *scikit-learn* (para *TF-IDF* e *Random Forest*), *Pytorch* (para implementação da *CNN-LSTM*), *Gensim* (para carregar embeddings *GloVe*) e as bibliotecas *TRL* e *Unsloth* (para o fine-tuning eficiente de *LLMs*).

Três abordagens distintas foram implementadas e avaliadas para a tarefa de classificação. A Figura 1 resume o processo utilizado.

A primeira aborgadem utilizou o *Random Forest* (RF), um modelo clássico de

Tabela 1. Estatísticas do Dataset: Nr. de Relatos por Natureza Ajustada

Natureza Ajustada	Quantidade de Relatos
Perda ou extravio de documento ou objeto	51.701
Estelionato	14.717
Furto	14.664
Acidente de trânsito sem vítima	8.945
Roubo	8.898
Outras Comunicações	5.117
Crime Cibernético	3.271
Ameaça	2.860
Dano	1.692
Difamação	1.063
Injúria	779
Calúnia	599
Perda ou extravio de documento e/ou objeto	541
Perturbação do sossego	504
Crimes Contra o Idoso	162
Injúria cometida ofendendo a dignidade ou o decoro	55
Perturbação do trabalho ou do sossego alheio	10
Violência Doméstica Contra Mulher	9
Acidente de trânsito sem feridos	1

aprendizado de máquina baseado em árvores de decisão. Os textos dos relatos passaram por um pré-processamento inicial (conversão para minúsculas, remoção de pontuações e *stop words* em português utilizando *NLTK*). Em seguida, foram vetorizados usando a técnica *TF-IDF* (*Term Frequency-Inverse Document Frequency*) com a implementação do *scikit-learn*. O modelo *Random Forest* foi treinado sobre esses vetores.

Outra abordagem explorada foi uma arquitetura de *deep learning* combinando Redes Neurais Convolucionais (CNN) e Redes Neurais Recorrentes do tipo *Long Short-Term Memory* (LSTM). Utilizou-se uma camada *Conv1D* (128 filtros, kernel de tamanho 5) para extrair características locais do texto, seguida por duas camadas LSTM (128 unidades, *dropout* de 0.5) para modelar dependências sequenciais. Como entrada, foram utilizados *word embeddings* pré-treinados *GloVe* (*Global Vectors*) de 1000 dimensões, específicos para o português do Brasil, disponibilizados pelo projeto Núcleo Interinstitucional de Linguística Computacional (NILC), que foram mantidos congelados durante o treinamento. O modelo foi implementado em *Pytorch*, utilizando otimizador *Adam* e um escalonador de taxa de aprendizado.

A terceira abordagem explorou o uso de um LLM pré-treinado, especificamente o modelo *Qwen 2.5* com 7 bilhões de parâmetros. O modelo foi adaptado para a tarefa de classificação através de *fine-tuning* utilizando a biblioteca Unsloth, que implementa otimizações para treinamento eficiente na GPU local, incluindo a técnica *QLoRA* (*Quantized Low-Rank Adaptation*) com quantização em 4 bits. A camada final de predição de tokens do LLM (`lm_head`) foi substituída por uma camada linear específica para as 10 classes do problema, forçando o modelo a gerar diretamente o rótulo da classe prevista.

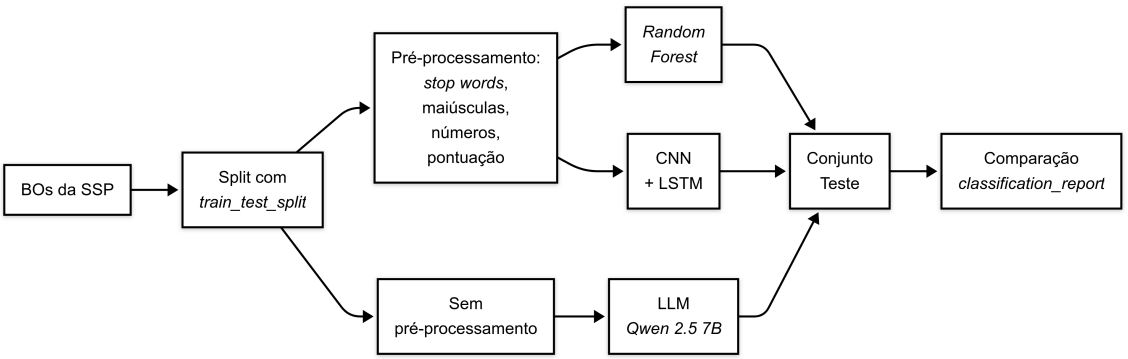

Figura 1. Fluxo de trabalho

4. Resultados Preliminares e Discussão

A análise exploratória inicial dos dados focou em 13 categorias selecionadas (após agrupamento e balanceamento). A análise da frequência de termos por categoria sugeriu a presença de padrões linguísticos distintos que poderiam ser explorados pelos modelos de classificação (Figuras 2, 3 e 4). No entanto, também evidenciou-se a necessidade de etapas de pré-processamento mais avançadas (como tratamento de acentos, lematização/*stemming*, substituição de valores ou símbolos específicos por um *token* especial) que não foram implementadas nesta fase inicial.

Os três modelos foram treinados no conjunto de treinamento balanceado e avaliados no conjunto de teste correspondente, utilizando a métrica *F1-Score* (*macro average*) como principal indicador de desempenho, por ser adequada para cenários multiclasse e fornecer uma visão balanceada entre precisão e *recall*. Os resultados obtidos, detalhados nas matrizes de confusão apresentadas, foram os seguintes: o *Random Forest* (com TF-IDF) alcançou *F1-Score* de 0.61 (Figura 5), a rede CNN-LSTM (com *GloVe* NILC) obteve *F1-Score* de 0.59 (Figura 6), e o modelo *Qwen 2.5 7B* (ajustado com QLoRA) atingiu *F1-Score* de 0.69 (Figura 7).

O *Qwen 2.5* com ajustes finos apresentou o melhor desempenho entre os modelos avaliados, alcançando um *F1-Score* de 0.69. Este resultado sugere que, mesmo com um ajuste fino relativamente limitado, a capacidade dos LLMs pré-treinados em compreender nuances da linguagem natural oferece vantagens sobre as abordagens clássicas e redes neurais específicas treinadas do zero ou com embeddings estáticos neste contexto.

A performance similar entre o *Random Forest* (0.61) e a CNN-LSTM (0.59) foi um resultado notável. Apesar das arquiteturas fundamentalmente diferentes, ambos os modelos apresentaram desempenho moderado e próximo. Isso levanta a hipótese de que a qualidade intrínseca dos dados, tais como: ruídos no texto, ambiguidades nos relatos, ou potenciais inconsistências na própria rotulagem da "Natureza Ajustada", pode atuar como um fator limitante mais significativo do que a capacidade representacional dos modelos em si.

A avaliação das classificações revelou dificuldades específicas, como a falta de clareza na definição da categoria "Outras comunicações". Adicionalmente, constatou-se que o modelo híbrido de redes neurais apresenta confusão entre as categorias "Crime Cibernético" e "Estelionato", o que provavelmente ocorre devido à similaridade entre as

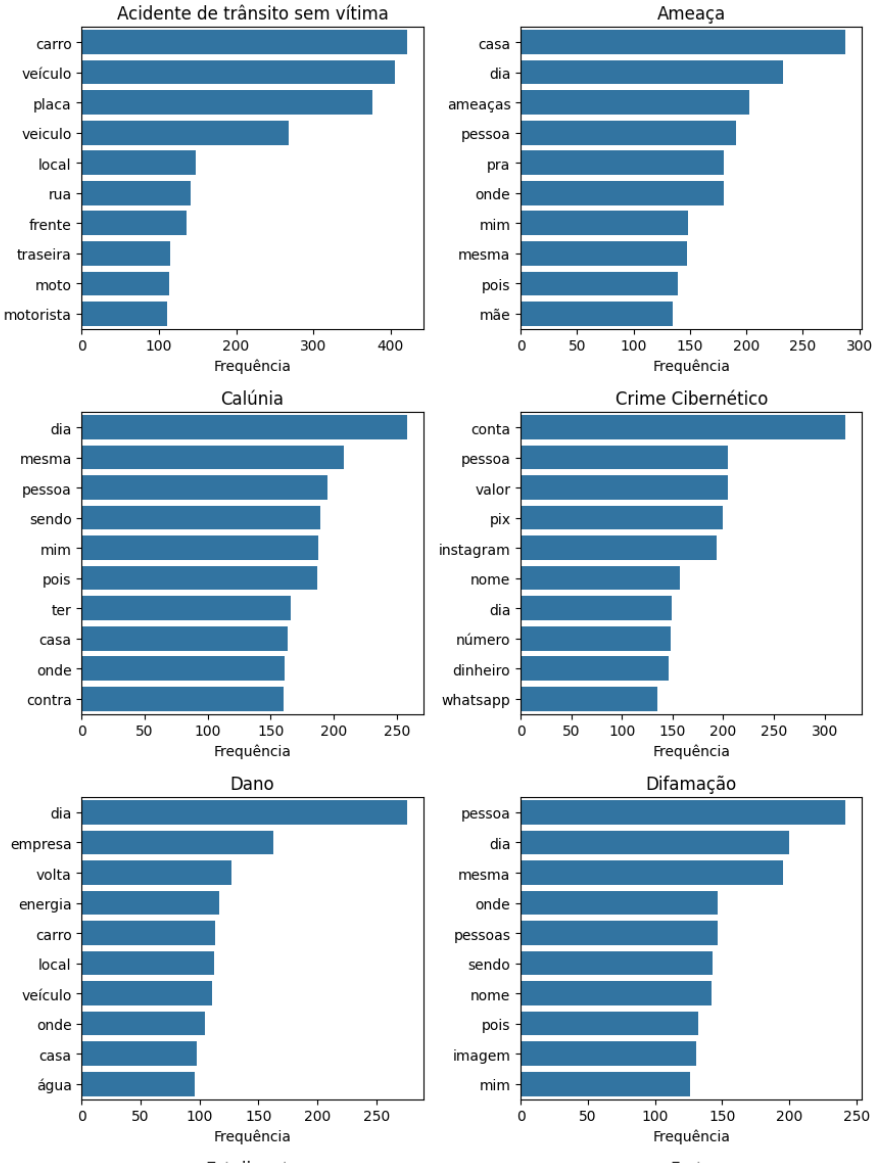

Figura 2. Palavras mais comuns por categoria. (Parte 1)

palavras mais frequentes encontradas nos relatos textuais de cada uma dessas classes.

Estes resultados preliminares, embora promissores especialmente para a abordagem com LLM, reforçam a importância de um pré-processamento de texto mais cuidadoso e, fundamentalmente, de uma análise aprofundada da qualidade e consistência dos rótulos do conjunto de dados.

5. Conclusão e Trabalhos Futuros

Este trabalho apresentou uma investigação preliminar sobre a aplicação de técnicas de PLN para a classificação automática de relatos textuais em Boletins de Ocorrência da SSP do Estado do Piauí. Os resultados indicam que Grandes Modelos de Linguagem (LLMs) fine-tuned, como o Qwen 2.5 7B, demonstram potencial superior às abordagens clássicas (Random Forest com TF-IDF) e redes neurais específicas (CNN-LSTM com GloVe) para esta tarefa, mesmo com limitações computacionais locais.

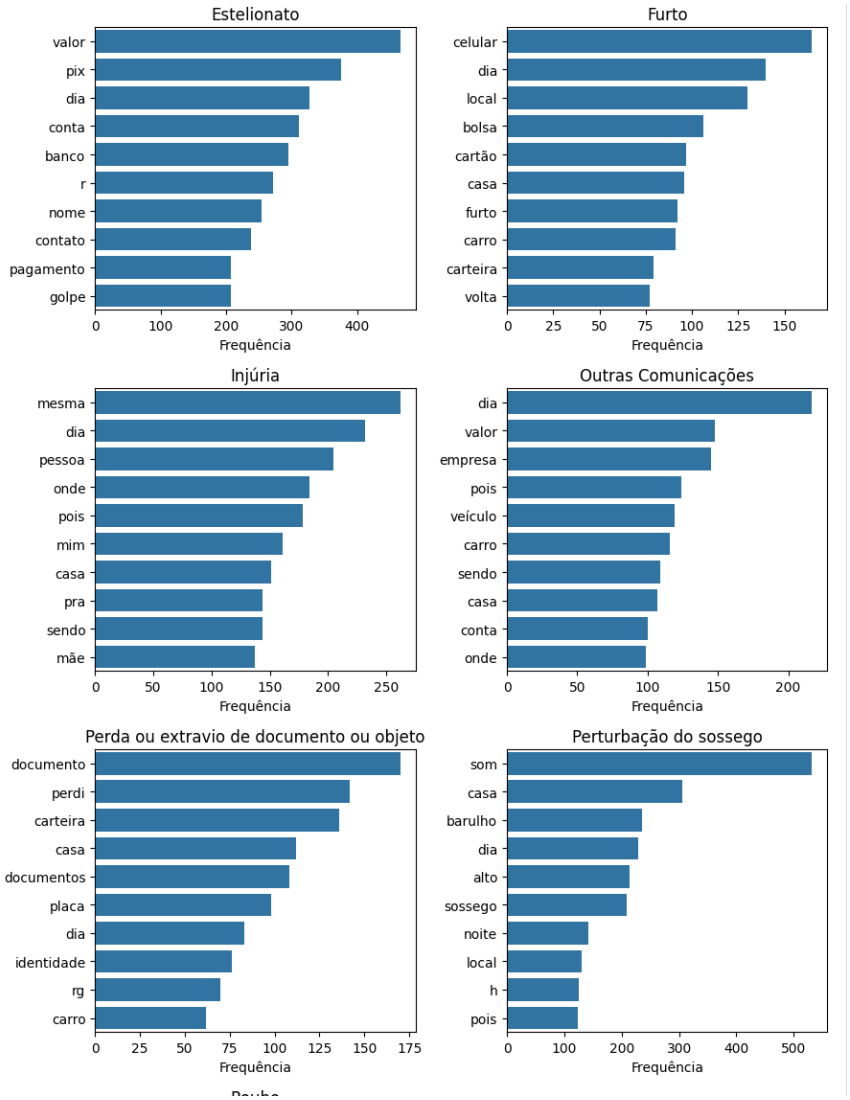

Figura 3. Palavras mais comuns por categoria. (Parte 2)

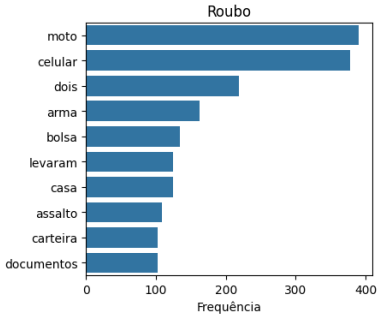

Figura 4. Palavras mais comuns por categoria. (Parte 3)

No entanto, a análise sugere que a qualidade dos dados textuais e dos rótulos associados é um fator crítico que limita o desempenho de todos os modelos avaliados. Diante

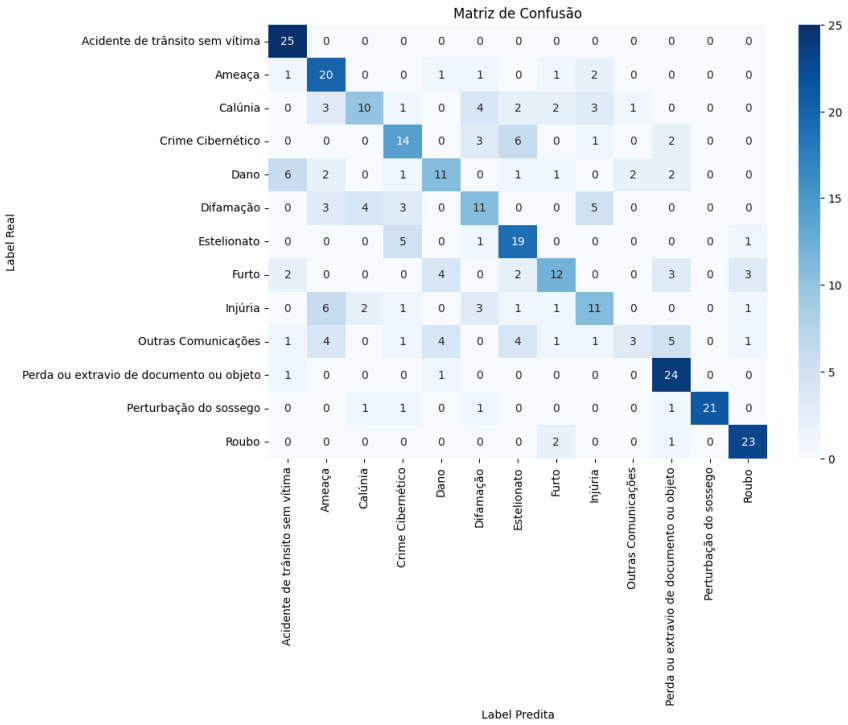

Figura 5. Matriz de Confusão - Random Forest (F1=0.61)

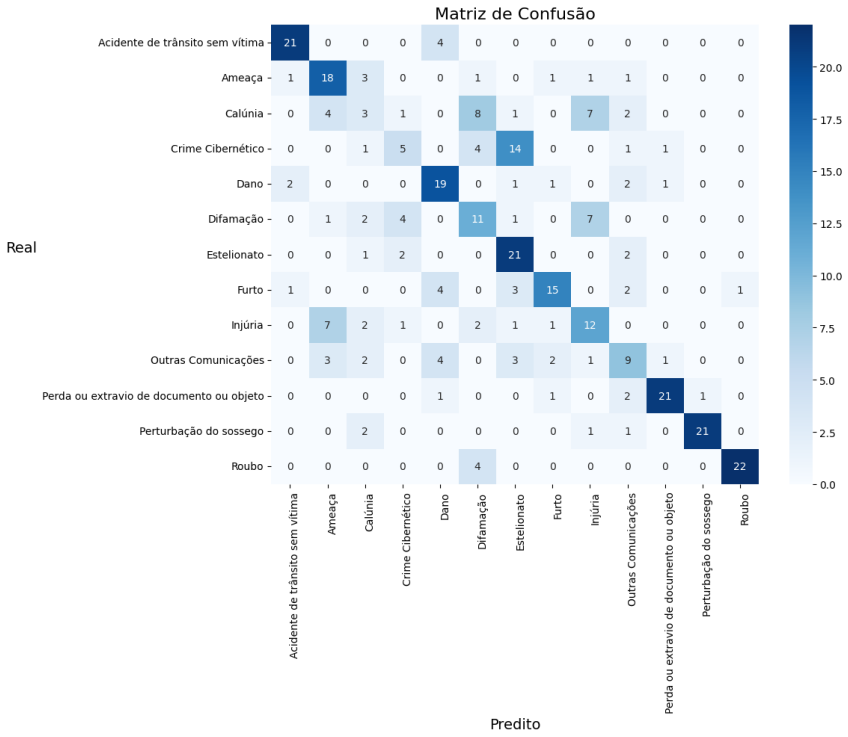

Figura 6. Matriz de Confusão - CNN-LSTM (F1=0.59)

disso, os próximos passos desta pesquisa concentrar-se-ão em duas frentes principais.

A primeira frente será a melhoria da qualidade dos dados. Isso envolverá

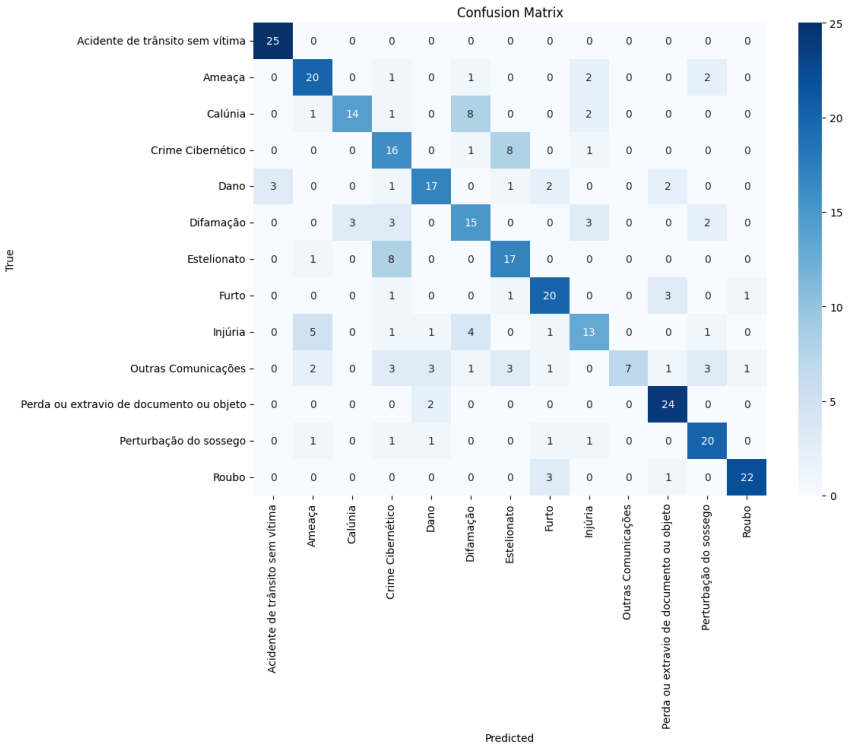

Figura 7. Matriz de Confusão - Qwen 2.5 7B (F1=0.69)

a implementação de rotinas de pré-processamento de texto mais sofisticadas, como lematização ou *stemming*, tratamento de entidades nomeadas e correção ortográfica. De forma crucial, planeja-se também realizar uma revisão e potencial re-rotulagem assistida de uma parte do conjunto de dados, visando aumentar a confiabilidade das classes atribuídas.

A segunda frente de trabalho será o foco na extração de informação. Uma vez que a etapa de classificação esteja mais robusta, ou mesmo utilizando a classificação atual como um filtro inicial, o objetivo será avançar para a extração de informações específicas contidas nos relatos – por exemplo, identificar objetos subtraídos, locais de ocorrência, características de suspeitos ou sequências de eventos. Para esta tarefa, pretende-se avaliar extensivamente a capacidade dos LLMs, como o próprio Qwen 2.5 ou modelos maiores caso os recursos disponíveis permitam. A abordagem incluirá a exploração de técnicas de engenharia de prompt (como *zero-shot* e *few-shot*) e, se for necessário, a realização de *fine-tuning* específico para a tarefa de extração, possivelmente empregando métodos como DPO (*Direct Preference Optimization*) ou GRPO (*Group Relative Policy Optimization*) para alinhar o comportamento do modelo aos requisitos da tarefa desejada.

Espera-se que a continuidade desta pesquisa contribua para o desenvolvimento de ferramentas mais eficazes para a análise de Boletins de Ocorrência (BOs), auxiliando assim o trabalho dos órgãos de segurança pública.

6. Agradecimentos

O presente trabalho foi financiado parcialmente pela Secretaria de Segurança Pública do Estado do Piauí (SSP/PI), com apoio da Fundação Cultural e de Fomento à Pesquisa, En-

sino, Extensão e Inovação (FADEX). Agradecemos também a equipe técnica da SSP/PI, Venceslau Felipe, Joaquim Carvalho e os demais colegas, pelo apoio e disponibilidade constante em conversar com os pesquisadores do Departamento de Computação da UFPI.

Referências

Amorim, M. S. and Pereira, J. R. S. (2019). Tipificação de ocorrências policiais utilizando machine learning. Trabalho de Conclusão de Curso (TCC).

Anjos Junior, O. and et al. (2020). Padrões de concentração espacial de roubos de automóveis em municípios da grande joão pessoa a partir de técnicas de aprendizado de máquinas. *Teoria e Prática em Administração*, 11(2):28–45.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Castro, U. R. M. (2020). Explorando aprendizagem supervisionada em dados heterogêneos para predição de crimes. Dissertação de mestrado, PUC-Minas.

Delegacia Virtual (2025). Delegacia virtual do ministério da justiça e segurança pública. Acesso em: 24 mar. 2025.

Kremer, G. R. (2023). Algoritmos de aprendizado de máquina aplicados a dados públicos para obtenção de insights em segurança pública. Trabalho de Conclusão de Curso (TCC).

Matos, H., Souza, S., Santos, R., Costa, J., and Costa, C. (2022). A supervised classifier for police reports at the state of pará, brazil. In *Anais da II Escola Regional de Alto Desempenho Norte 2 e II Escola Regional de Aprendizado de Máquina e Inteligência Artificial Norte 2*, pages 21–24, Porto Alegre, RS, Brasil. SBC.