Classificadores Supervisionados para Relatos em Boletins de Ocorrências Policiais

Carolina Santos Lins, Filipe Saraiva

[1]Faculdade de Computação – Instituto de Ciências Exatas e Naturais
Universidade Federal do Pará (UFPA)
Av. Augusto Correa 01, 66075-090 – Belém – PA – Brasil

carolina.lins@icen.ufpa.br, saraiva@ufpa.br

Abstract. *Due to the significant incidence of property crimes consistently recorded in police reports, databases of public security accumulate a substantial number of such cases. One of the responsibilities of the State of Pará's Sub-Secretariat of Intelligence and Criminal Analysis, an entity subordinate to the Secretary of Public Security and Social Defense of the State of Pará, is the classification of these occurrences, aiming not only for statistical purposes but also to provide support for decision-making related to public security. This work describes the process of developing supervised classifiers, encompassing the methodology employed to handle extremely imbalanced data, a common characteristic in datasets related to criminal incidents. Three distinct classifiers were built: one to identify the location, another to classify the victim, and the last one to identify the items and the modus operandi. They achieved an accuracy of 75,86% 90,33% and 80,56% respectively.*

Resumo. *Devido à considerável incidência de crimes contra o patrimônio constantemente registrados nos boletins de ocorrência policiais, as bases de dados de segurança pública acumulam uma grande quantidade desses casos. Uma das responsabilidades da Secretaria Adjunta de Inteligência e Análise Criminal do Estado do Pará, órgão subordinado à Secretaria de Estado de Segurança Pública e Defesa Social do Pará, é a classificação dessas ocorrências, visando não apenas fins estatísticos, mas também fornecendo apoio para a tomada de decisões relacionadas à segurança pública. Este trabalho descreve o processo de desenvolvimento de classificadores supervisionados, abrangendo a metodologia empregada para lidar com dados extremamente desbalanceados, uma característica comum em conjuntos de dados relacionados a ocorrências criminais. Foram construídos três classificadores distintos: um para identificar a localidade, outro para classificar a vítima e o último para identificar os itens e o modus operandi. Eles alcançaram uma acurácia de 75,86%, 90,33% e 80,56% respectivamente.*

1. Introdução

No âmbito da segurança pública, os crimes contra o patrimônio são consistentemente reportados através de boletins de ocorrência policiais e representam uma preocupação central para as autoridades. Dentro dessa categoria, destacam-se o furto e o roubo como algumas das ocorrências mais comuns, que anualmente registram centenas de milhares de ocorrências no Brasil [Fórum Brasileiro de Segurança Pública 2023].

Portanto, os órgãos públicos encarregados de compilar informações estatísticas sobre segurança pública mantêm extensas bases de dados contendo milhares de registros detalhados relacionados a esse tipo de ocorrência. Essas informações, preservadas pelos órgãos competentes, desempenham um papel crucial ao serem empregadas em pesquisas, análises e na elaboração de medidas efetivas no combate aos crimes.

Contudo, não apenas a informação bruta é preservada. As ocorrências registradas em delegacias já contêm uma quantidade substancial de informações sobre os fatos; no entanto, elas passam por uma análise minuciosa para extrair novos dados que serão incorporados às bases de dados, visando manter as informações estatísticas o mais detalhadas possível.

No estado do Pará, a Secretaria Adjunta de Inteligência e Análise Criminal (SIAC), subordinada à Secretaria de Segurança Pública e Defesa Social, é responsável por centralizar, consolidar e disponibilizar os dados estatísticos oficiais relativos à segurança pública [SEGUP-PA 2023]. Uma das várias atribuições dos profissionais na SIAC é a categorização dos eventos registrados nos boletins de ocorrência policial. Esses boletins são formalizados nas delegacias e, posteriormente, essas informações são arquivadas em bases de dados acessíveis na própria Secretaria. Muitos detalhes já são inseridos no momento do registro da ocorrência, sendo o relato o campo principal, notadamente aquele que, na maioria dos casos, apresenta a maior riqueza de informações sobre o evento. É precisamente desse campo que os dados serão extraídos para efetuar a classificação e enriquecer ainda mais as informações relacionadas à ocorrência dentro do banco de dados da SIAC.

Um extenso trabalho de classificação de ocorrências é conduzido na SIAC. Nessa tarefa, vários campos são preenchidos na base de dados do órgão, incluindo um campo dedicado à especificação do crime. Este campo é destinado a informações adicionais sobre o fato relatado, mas não necessariamente indica a tipificação penal. No contexto de crimes patrimoniais, como furto e roubo, a especificação abrange detalhes como o *modus operandi*, o local, o tipo de vítima (pessoa física ou jurídica) e os itens subtraídos. A Figura 1 exemplifica como a especificação seria realizada a partir dos relatos fictícios apresentados.

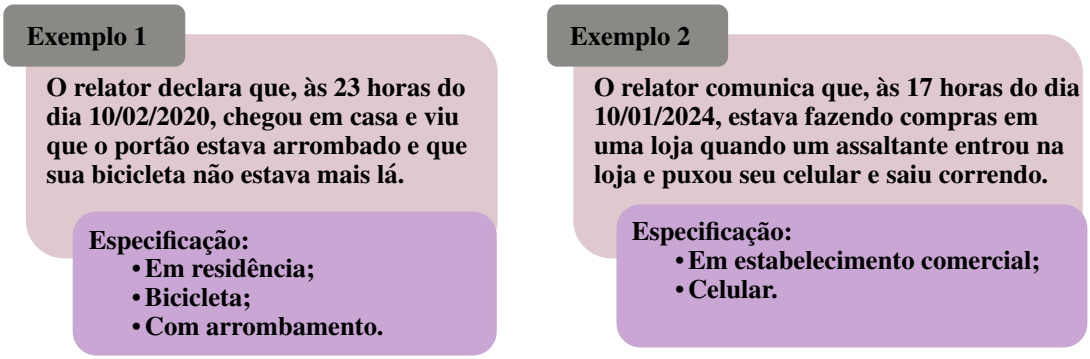

Figura 1. Exemplos de especificação a partir de relatos.

O objetivo central do presente estudo é a concepção e a implementação de um

classificador supervisionado, cuja finalidade principal seja realizar uma análise automatizada do conteúdo narrativo presente nos boletins de ocorrências relacionados a casos de roubo e furto. Esta iniciativa surge da necessidade de aprimorar e otimizar os processos de interpretação das informações detalhadas nessas ocorrências.

Este artigo está estruturado da seguinte forma: a **Seção 2** explora trabalhos correlatos; a **Seção 3** apresenta minuciosamente as etapas de pré-processamento de dados; a **Seção 4** apresenta a justificativa para a utilização de múltiplos classificadores e as técnicas empregadas para lidar com o desbalanceamento; a **Seção 5** traz os resultados dos classificadores individualmente e em conjunto; por fim, a **Seção 6** recapitula os principais pontos e define possíveis direções para pesquisas futuras.

2. Trabalhos Relacionados

Foram identificados vários estudos na área de inteligência computacional voltados para análise de descrições textuais de crimes.

No trabalho conduzido por [dos Reis Matos 2022], foi desenvolvido um classificador supervisionado que faz uso das descrições de eventos constantes nos boletins de ocorrência policial para categorizar entre 463 classes relacionadas à segurança pública. Os dados utilizados para realizar o estudo são provenientes da Secretaria Adjunta de Inteligência e Análise Criminal, no estado do Pará. O classificador, que consiste em uma rede neural com três camadas convolucionais, obteve aproximadamente 78% de acurácia. A implementação resultou em melhorias significativas no processo de classificação realizado pelos analistas criminais na SIAC.

O estudo realizado por [Oliveira 2020] explora a utilização de técnicas de Aprendizado Profundo, especificamente em Reconhecimento de Entidade Nomeada, para automatizar a extração de informações relevantes de narrativas de crimes contra o patrimônio. Tais narrativas são provenientes da base de dados da Secretaria de Segurança Pública e Defesa Social do Estado do Ceará. Como parte do âmbito do trabalho, questões relacionadas ao desbalanceamento de dados e representação vetorial eficiente do vocabulário foram abordadas. Sendo que diferentes tipos de *word embeddings* foram combinadas em um modelo *BiLSTM-CRF* (composto por uma camada *LSTM* bidirecional e uma camada *CRF*). Quanto aos resultados, os melhores modelos obtiveram 88,1% de acurácia balanceada.

No artigo de [Qi 2020], o escopo da pesquisa aborda a classificação de diferentes tipos de crimes de furto que ocorreram em uma cidade no período de 2009 a 2019. A classificação é feita a partir de dados textuais e o pré-processamento é feito por meio da técnica *TF-IDF*. O trabalho concentra-se no algoritmo *XGBoost*, que consiste na combinação de múltiplas árvores de decisão por meio de *boosting*. A performance deste método é comparada com a de diversos outros tipos de algoritmos como *KNN*, *SVM* e *Naive Bayes*.

No trabalho de [Li et al. 2020], a extração de eventos em textos jurídicos chineses é explorada. O escopo do trabalho está relacionado a casos de furto, e alguns dos eventos definidos no artigo são ato do furto, gastar dinheiro roubado, sacar e transferir dinheiro por meio de contas bancárias ou cartões roubados. No que diz respeito à estratégia adotada, a obtenção dos vetores foi realizada utilizando *BERT*, enquanto a extração foi feita por meio dos modelos *BiLSTM-CRF* e *CRF*.

No estudo [Kuang et al. 2017], a pesquisa foca na busca por classes que representem descrições textuais de crimes, indo além das categorias definidas pela classificação legal. Quanto à metodologia, a abordagem adotada inclui o método *TF-IDF* para vetorização e a técnica chamada *Non-Negative Matrix Factorization* (*NMF*) para detecção de tópicos. Assim, torna-se viável obter informações adicionais relacionadas a determinados tipos de crimes, possibilitando análises mais aprofundadas.

No artigo [Birks et al. 2020], uma pesquisa é realizada para identificar diversos *modus operandi* a partir de descrições textuais de crimes de roubo em residências, utilizando técnicas de aprendizado não supervisionado. Quanto ao método, o texto é representado através de *bag of words*, e o modelo estatístico conhecido como Alocação Latente de Dirichlet é utilizado para a modelagem de tópicos.

Em [Das et al. 2019], é utilizado técnicas de processamento de linguagem natural para extrair relações entre entidades nomeadas de descrições textuais de crimes contra mulheres na Índia, tais dados são coletados de notícias sobre esses crimes. Os pesquisadores propuseram um algoritmo de agrupamento hierárquico baseado em grafo para extrair relações entre as entidades nomeadas.

3. Dados e Pré-processamento

Esta seção fornece detalhes sobre a origem dos dados, sua quantidade e a divisão para treinamento e teste. Além disso, são exploradas em detalhes todas as etapas de pré-processamento, abrangendo o tratamento das classes e dos textos descritivos utilizados como entrada para os classificadores.

3.1. Dados Utilizados

Para desenvolver os classificadores, foram coletados aproximadamente 630 mil registros de furtos e roubos extraídos da base da SIAC. No órgão, a especificação mais detalhada dos relatos de crimes de roubo começou somente a partir de 2019, e para os crimes de furto, a partir de 2020. Consequentemente, os conjuntos de dados utilizados para treinamento incluíram o período de 2019 a 2022 para ocorrências de roubo e de 2020 a 2022 para furtos. As ocorrências relativas ao ano de 2023, mais especificamente de janeiro até julho de 2023, foram reservadas para a fase de teste. Essa divisão resulta em aproximadamente 86% dos dados destinados ao treinamento e 14% dos dados destinados para teste, em relação ao total dos dados.

3.2. Pré-processamento

3.2.1. Ajuste da Especificação

Na SIAC, os leitores que avaliam a ocorrência são responsáveis por digitar a especificação do crime, a qual é o campo dentro da base de dados do Órgão, onde são registradas as informações adicionais sobre o crime. Como mencionado anteriormente, a especificação dos crimes de furtos e roubos é a mais detalhada dentro do órgão, abrangendo detalhes como o local do incidente, o *modus operandi* e os itens roubados, entre outros. Todas essas informações são registradas neste mesmo campo, mas são separadas por barra. Dado que esse campo é digitado, é comum ocorrerem erros de digitação.

Para extrair as classes associadas a cada relato, o campo textual correspondente à especificação foi fragmentado a cada barra presente no texto, formando assim listas de

classes distintas. A fim de corrigir os equívocos existentes neste campo, todas as classes identificadas foram verificadas e a quantidade de ocorrências de cada uma foi contabilizada. Observou-se frequentemente que, na lista de classes construída inicialmente, aquelas que contabilizavam poucas dezenas de ocorrências eram, em sua maioria, classes majoritárias mas escritas com erros de digitação, sendo esses prontamente corrigidos.

Contudo, mesmo após essa correção, a lista de classes ainda abrigava diversas classes com um número reduzido de ocorrências ao longo dos anos de boletins registrados na base de dados. Nessa fase, com a colaboração de especialistas responsáveis pela classificação, algumas classes foram removidas, enquanto outras foram consolidadas com outras classes afins. Essa abordagem visava aprimorar a representatividade das classes consideradas no processo de classificação. Ao término desta fase, das 323 classificações potenciais que estavam inicialmente consideradas, restaram apenas 130 classes. Esse processo resultou em uma redução significativa, concentrando e simplificando a estrutura de classificação para um conjunto mais gerenciável e representativo.

3.2.2. Eliminação de dados incompletos

Para melhorar a qualidade dos dados, o processo de preparação dos dados iniciou-se com a eliminação de registros que não continham informações na especificação do crime, assim como os registros com relatos duplicados. Com isto, sobram 597962 linhas da base de dados coletada.

Adicionalmente, durante a fase de ajuste da especificação, foram identificadas as classes a serem removidas. Todos os registros na base de treino que continham alguma dessas classes na especificação foram excluídos, mesmo que apresentassem outras classes, pois a presença de uma classe inválida sugere uma maior probabilidade de outros possíveis erros de classificação. Após este filtro, a base contém apenas 546461 registros.

Além disso, uma quantidade considerável de registros continha apenas a especificação “OUTROS” ou “OUTRAS ESPECIFICAÇÕES”, indicando casos de furto ou roubo de objetos que não se enquadram em nenhuma das classes especificadas na SIAC. Contudo, devido a mudanças ao longo dos anos na metodologia de classificação, em que objetos antes não classificados passaram a ter classes específicas, optou-se por remover as ocorrências com apenas essas especificações. Por fim, o tamanho final da base utilizada é de 530894 linhas.

3.2.3. Tratamento dos dados descritivos do fato

Com o intuito de utilizar o relato presente no boletim de ocorrência como entrada para a rede neural, uma série de ajustes foi executada.

Inicialmente, todo o conteúdo foi convertido para letras minúsculas, uma vez que não havia padronização nos dados coletados nesse aspecto. Ainda, em várias ocasiões, o conteúdo textual que descreve o evento na base de dados é extraído diretamente do boletim de ocorrência, conservando diversas marcas HTML no texto. Consequentemente, foi feita a exclusão de todos os tipos de *tags* HTML do texto. Por fim, foram implementadas correções adicionais, incluindo a remoção de caracteres repetidos que surgiam no início

ou no final do conteúdo textual, bem como a eliminação de espaços duplicados.

Essas etapas iniciais foram conduzidas no projeto de [dos Reis Matos 2022], que descreve um trabalho realizado a partir de relatos de ocorrências provenientes da mesma base utilizada nesta pesquisa.

Com isso, sobra somente o conteúdo textual com significação sobre o fato narrado no boletim. Ainda assim, havia alguns elementos textuais muito frequentes nos relatos, mas que não continham informações relevantes para a classificação. Decidiu-se portanto retirar os cabeçalhos padronizados, que ocorriam em grande parte das ocorrências, principalmente aquelas registradas por meio da Delegacia Virtual. Tal remoção foi feita utilizando expressões regulares.

Melhorar o desempenho da rede neural ao processar texto frequentemente envolve a normalização e, nesse contexto, a lematização destaca-se como uma etapa essencial. Essa técnica consiste em mapear as palavras para sua forma raiz [Jurafsky and Martin 2009]. Algumas ferramentas capazes de realizar este processo em dados em português estão disponíveis, dentre elas o módulo *spacy* [Honnibal et al. 2020] destaca-se como uma das principais opções, pois é multifuncional no contexto de tratamento de texto para aprendizado de máquina [Caseli and Nunes 2023]. Esse módulo foi empregado na execução da lematização e em outras fases do processamento textual.

Após a conclusão da lematização dos dados, o próximo passo envolveu a remoção das denominadas "palavras de parada" do texto. Estas são termos que, embora apareçam com frequência, possuem escasso significado para a classificação. Para efetuar essa exclusão, foi elaborada uma lista de *stopwords* que seria aplicada nos dados textuais já submetidos à lematização. Essa relação teve como base as "palavras de parada" presentes nas bibliotecas *nltk* e *spacy*. No entanto, uma consideração especial foi dada à preservação de palavras que indicam negação. Essa decisão foi motivada pelo fato de que a especificação do crime na Secretaria abrange uma gama de detalhes da ocorrência, e a presença ou ausência dessas palavras pode exercer uma influência substancial no significado atribuído ao texto. Por exemplo, no seguinte relato fictício, a vítima declara: "percebi a ausência de diversos itens de valor na minha casa depois que voltei da viagem, mas não notei sinais de arrombamento", a remoção dos termos que indicam negação alterariam o sentido e levariam à classificação errônea referente à classe "COM ARROMBAMENTO".

Adicionalmente a esses termos, procedeu-se à eliminação dos numerais, uma vez que, frequentemente, esses elementos denotam informações pessoais do relator, como CPF e número de telefone. Tais detalhes possuem pouquíssima relevância para o processo de classificação.

Finalizando, após a conclusão de todo o pré-processamento textual e a tokenização realizada a nível de palavras, optou-se por utilizar a metodologia de *word embeddings*. Essa escolha é fundamentada por diversos motivos, sendo o primeiro a crucial importância da ordem das palavras para uma classificação precisa, elemento que é perdido em métodos como o *bag of words* [Abubakar et al. 2022]. Além disso, o segundo motivo diz respeito à extensão considerável do vocabulário, tornando o uso de *embeddings* uma alternativa mais eficiente em comparação com abordagens que lidam diretamente com o vocabulário completo [Jurafsky and Martin 2009].

4. Metodologia

Esta seção é organizada da seguinte maneira: a Seção 4.1 aborda a motivação por trás da utilização de múltiplos classificadores; a Seção 4.2 detalhas as métricas que serão utilizadas para a análise de resultados; a Seção 4.3 descreve a construção dos modelos e arquitetura proposta para cada um.

4.1. Abordagem de Múltiplos Classificadores

Optou-se por abordar o desafio da classificação em duas fases distintas. Inicialmente, a decisão foi de tratar o problema como uma tarefa de classificação multirrótulo, uma vez que um registro pode pertencer a várias classes simultaneamente. Contudo, logo se percebeu a presença de um conjunto de classes que não poderiam ser positivas ao mesmo tempo para nenhum dos registros. Essas classes referem-se à localização e a estratégia adotada foi dividir o problema. As classes relacionadas a *modus operandi*, objetos roubados e outras especificações foram direcionadas para um classificador multirrótulo, o qual é capaz de associar múltiplas classes a uma única instância. Enquanto as classes associadas à localização foram destinadas a um classificador multiclasse e monorrótulo, o qual classifica uma instância em uma única classe das múltiplas classes possíveis.

Além disso, estratégia de múltiplos classificadores viabiliza a abordagem individual de cada grupo de categorias na classificação, possibilitando a implementação de soluções específicas para o desbalanceamento sem interferir nas demais classes em classificadores distintos. Essa abordagem visa otimizar a precisão e eficácia do processo de classificação em relação à natureza específica das classes envolvidas.

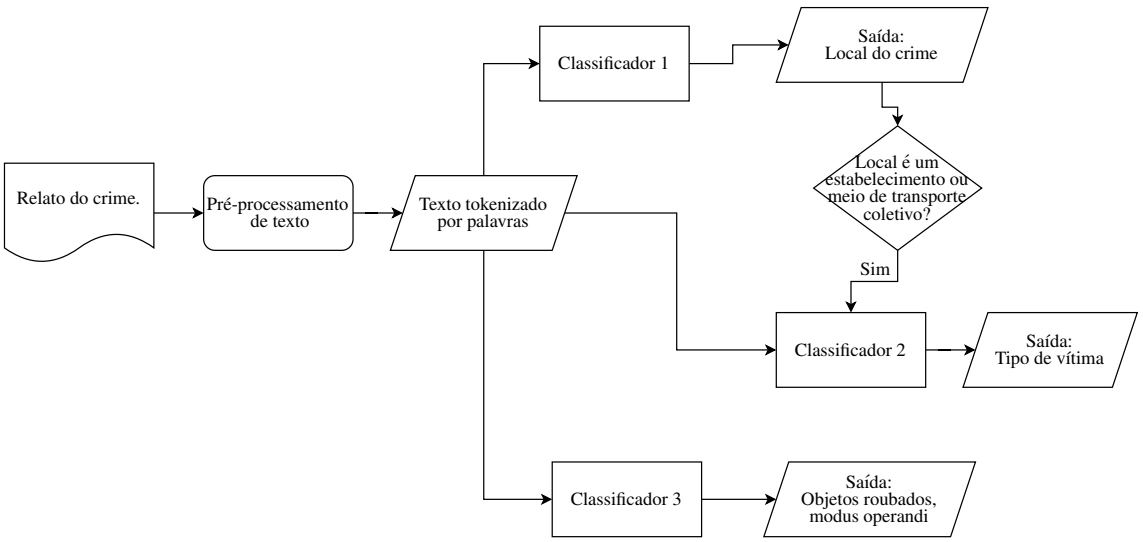

Figura 2. Fluxo do sistema.

A Figura 2 ilustra o fluxo do sistema com a abordagem de utilizar múltiplos modelos, destacando a interação entre os classificadores 1 e 2 e as informações que cada um deles busca classificar.

4.1.1. Classificação quanto a Localidade

O primeiro classificador desenvolvido concentrou-se na atribuição de localização. Inicialmente, existiam 23 classes indicativas de locais, no entanto, algumas dessas classes representavam o mesmo local, mas com detalhes específicos de classificação adotados pela metodologia da Secretaria. Como resultado, essas classes foram aglutinadas, totalizando assim 15 classes de localização.

A Figura 3 ilustra a proporção das classes referentes à localização nos dados de treinamento para a rede neural, após as etapas de pré-processamento detalhadas na Seção 3.2.

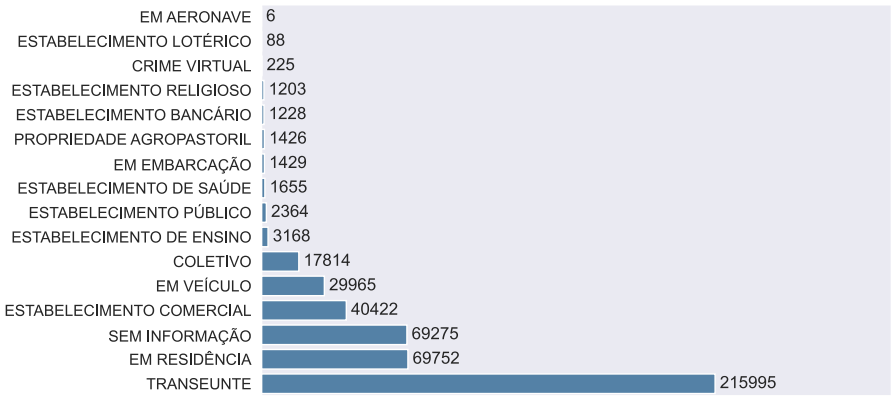

Figura 3. Proporção das classes indicativas de local nos dados de treino.

Na Figura 3 apresentada, nota-se um desequilíbrio significativo entre essas classes, com "TRANSEUNTE" destacando-se como a predominante. Essa classe indica que o crime teve como alvo uma pessoa que estava em via pública ou logradouro público.

Outra ponto a se destacar na Figura 3 é a grande quantidade de instâncias que não contêm informações sobre a localização do crime. Porém, é válido manter uma classe que indica a ausência de informações sobre a localidade, uma vez que, de acordo com a metodologia utilizada na SIAC, existem casos nos quais não há a determinação do tipo de local. A preservação dessa classificação é fundamental para evitar uma disparidade substancial entre as atribuições de classe realizadas pelos leitores de relatos na SIAC e aquelas feitas pela rede neural. Isso contribui para a coerência e alinhamento entre os métodos de classificação adotados.

4.1.2. Classificação referente à Vítima

Na metodologia empregada pela SIAC, um dos aspectos de grande importância durante a análise dos relatos é a identificação da vítima do crime. Embora existam campos específicos na base de dados para registrar informações das vítimas, é comum encontrar, na especificação do crime, detalhes sobre a vítima, indicando se é uma pessoa física ou jurídica. Essa ocorrência esta presente nos casos em que o crime se desenrola em um estabelecimento ou meio de transporte coletivo. Nestas situações, a especificação do crime informa se o roubo ou furto ocorreu apenas nesses locais ou se envolveu a apropriação de

propriedades do local.

Para ilustrar, se um crime de roubo acontece dentro de um estabelecimento comercial e a vítima é uma pessoa física, a classificação seria “em estabelecimento comercial”. Contudo, se as propriedades do estabelecimento são alvo do roubo, a classificação passa a ser “ao estabelecimento comercial”.

Este tipo de classificação, que distingue se a vítima é uma pessoa física ou jurídica, só é necessária nos casos específicos em que a localidade é um estabelecimento ou meio de transporte coletivo. Portanto, o classificador para este problema de classificação foi treinado apenas com uma parte da base de treinamento correspondente a esses casos específicos e é utilizado somente quando o classificador de locais identifica essas situações.

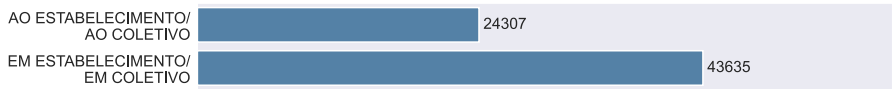

Figura 4. Proporção das classes indicativas do tipo de vítima nos dados de treino.

Como é possível visualizar na Figura 4, dentre todos os classificadores detalhados neste trabalho, este é o único que não enfrenta o desafio do desbalanceamento extremo de classes.

4.1.3. Classificação Multirrótulo

O último classificador desenvolvido abrange as demais 107 classes da especificação do crime, englobando *modus operandi*, objetos furtados ou roubados, e informações adicionais sobre a vítima. A distribuição das classes mais frequentes nos dados de treino está apresentada na Figura 5. Notavelmente, ao contrário da classificação de localidade, esta categorização permite a atribuição de mais de uma classe para um determinado registro, pois pode haver diversas combinações de *modus operandi* e vários objetos roubados. Nesse sentido, o classificador desenvolvido para este problema é multirrótulo.

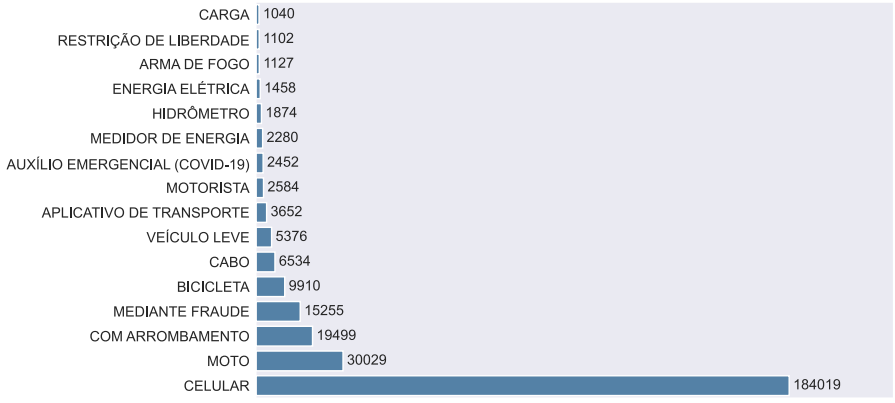

Figura 5. Proporção das classes mais frequentes do classificador multirrótulo nos dados de treino.

Na Figura 5, é notável que a classe mais comum associada ao item subtraído é "CELULAR", enquanto a classe mais frequente relacionada ao *modus operandi* é "COM ARROMBAMENTO".

4.2. Métricas para Medir a Performance

As situações aqui apresentadas envolvem diferentes tipos de problemas classificação sendo:

- Classificador 1: Problema de classificação multiclasse com desbalanceamento de dados.
- Classificador 2: Problema de classificação binária.
- Classificador 3: Problema de classificação multirrótulo com desbalanceamento de dados.

Considerando isto, as métricas que serão aplicadas para indicar a performance serão:

- **Acurácia:** A acurácia é definida como a razão entre o número de predições corretas e o número total de predições. É representada pela seguinte expressão:

$$\text{Acurácia} = \frac{\text{Número de Predições Corretas}}{\text{Número Total de Predições}}$$

- **Precisão:** A precisão é a razão entre o número de verdadeiros positivos e a soma dos verdadeiros positivos e falsos positivos. É representada pela seguinte expressão:

$$\text{Precisão} = \frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos + Falsos Positivos}}$$

- **Revocação (Sensibilidade):** A revocação é a razão entre o número de verdadeiros positivos e a soma dos verdadeiros positivos e falsos negativos. É representada pela seguinte expressão:

$$\text{Revocação} = \frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos + Falsos Negativos}}$$

- ***F1-score*:** O *F1-score* é a média harmônica entre precisão e revocação. É representada pela seguinte expressão:

$$\textit{F1-score} = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão + Revocação}}$$

- **Acurácia Balanceada:** A acurácia balanceada leva em consideração o desbalanceamento de classes e é calculada como a média das sensibilidades (revocação) de cada classe. É representada pela seguinte expressão:

$$\text{Acurácia Balanceada} = \frac{1}{n} \sum_{i=1}^{n} \text{Revocação}_i$$

- ***Hamming Loss*:** A *Hamming Loss* representa a fração média de rótulos erroneamente preditos. É representada pela seguinte expressão:

$$\textit{Hamming Loss} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|L_i|}\sum_{j=1}^{|L_i|}\text{XOR}(\text{Predição}_{ij}, \text{Rótulo}_{ij})$$

- **Coeficiente de similaridade de Jaccard:** O coeficiente de similaridade de Jaccard, também conhecido como Índice de Jaccard, mede a similaridade entre conjuntos e será usado para avaliar a sobreposição entre a predição e o rótulo. É representada pela seguinte expressão, onde *A* e *B* são conjuntos [Leskovec et al. 2014]:

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

4.3. Construção dos Modelos

Esta seção descreve detalhadamente a arquitetura de cada classificador e as estratégias utilizadas. As arquiteturas foram encontradas de forma empírica.

Adicionalmente, todos os classificadores descritos neste trabalho foram implementados utilizando os módulos *TensorFlow* [Abadi et al. 2015] e *Keras* [Chollet et al. 2015].

4.3.1. Classificador de Localidade

Quando se trata de tarefas de Processamento de Linguagem Natural (PLN), uma das principais arquiteturas de redes neurais amplamente empregadas nesses cenários é a *Long Short-Term Memory* (*LSTM*), a qual têm capacidade de capturar dependências de longo prazo, o que a torna especialmente adequada para lidar com a natureza sequencial e contextual dos dados de texto [Patel and Arasanipalai 2021].

As redes neurais convolucionais, embora amplamente aplicadas em tarefas de visão computacional, também são aplicadas em tarefas de Processamento de Linguagem Natural (PLN) [Patel and Arasanipalai 2021]. Este cenário é evidenciado no estudo detalhado por [dos Reis Matos 2022], onde uma rede neural convolucional obteve aproximadamente 78% de acurácia na classificação de relatos de boletins de ocorrência.

Vários estudos, como [Wang et al. 2020], [Luan and Lin 2019] e [Li and Ning 2020], destacam as vantagens da combinação dessas duas arquiteturas em tarefas de classificação textual, unindo a capacidade de compreensão contextual oferecida por *LSTMs* com a eficácia na extração de características por meio de filtros proporcionada por *CNNs* (*Convolutional Neural Networks*).

Assim, a estrutura adotada para esta tarefa específica de classificação compreende um modelo com as camadas a seguir:

- **Camada de Entrada:** recebe sequências numéricas de comprimento 512, onde os números correspondem à palavras.
- **Camada de *Embedding*:** responsável por transformar a entrada em vetores densos, permitindo a representação vetorizada das palavras.

- **Camada de *Spatial Dropout*:** utilizada para normalização e prevenção de *overfitting*, descartando aleatoriamente valores durante o treinamento para melhorar a generalização do modelo.
- **Camada *LSTM*:** captura relações sequenciais de longo alcance.
- **Camada Convolucional:** extrai características por meio da aplicação de filtros, identificando padrões relevantes em diferentes regiões do *input*.
- **Camada *Global Max Pooling*:** realiza o processo de *pooling* global máximo, esta camada reduz a dimensionalidade do tensor resultante da camada anterior, extraindo os valores máximos.
- **Camada de Saída Densa:** camada final que produz a saída do modelo, aplicando a função *softmax* e gerando probabilidades normalizadas.

A Figura 6 ilustra a arquitetura deste modelo. No módulo *Keras*, o termo "None"é utilizado para indicar dimensões variáveis ou flexíveis em tensores. Quando o primeiro valor nas dimensões de entrada e de saída é "None", isso representa a possibilidade de lidar com lotes de diferentes tamanhos durante o treinamento e a inferência do modelo.

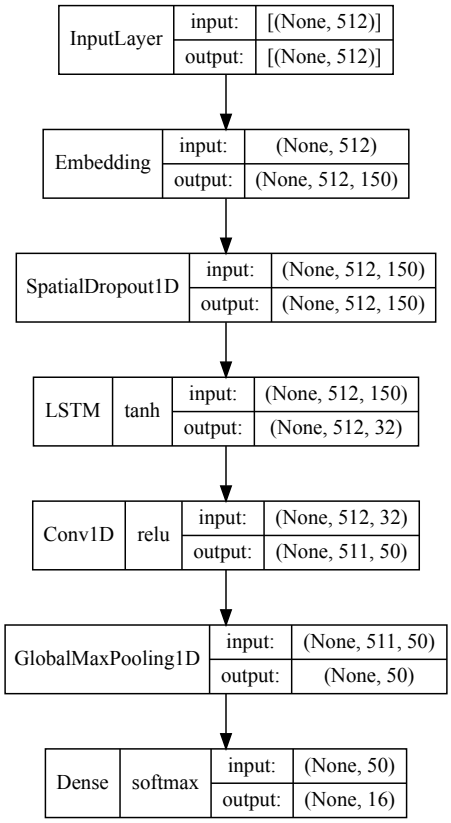

Figura 6. Modelo da rede neural para a classificação de localidades.

Adicionalmente, na fase de treinamento, optou-se por empregar a função de perda *categorical crossentropy*, incorporando o *label smoothing* como método adicional de regularização. Essa estratégia visa mitigar o impacto de predições excessivamente confiantes, especialmente considerando a quantidade de erros de classificação nos dados. O *label smoothing* consiste na suavização das probabilidades atribuídas às classes, isto diminui o impacto de dados errôneos no treinamento [Goodfellow et al. 2016].

O principal desafio nesta tarefa de classificação reside no desbalanceamento extremo entre as classes. Um método para lidar com esse desbalanceamento é o balanceamento de pesos associados a cada classe [Johnson and Khoshgoftaar 2019]. Essa estratégia possibilita que o classificador dê atenção às classes minoritárias sem recorrer necessariamente a métodos de *undersampling*, nos quais informações valiosas podem ser descartadas, ou *oversampling*, que tende a gerar *overfitting* [Maimon and Rokach 2010].

4.3.2. Classificador de Vítima

Dado que a tarefa de classificação do tipo de vítima é menos complexa em comparação com outras, optou-se por desenvolver um modelo de dimensões relativamente menores em comparação com outros classificadores. Além disso, observando que arquiteturas convolucionais de forma geral são notavelmente rápidas e capazes de lidar com tarefas de Processamento de Linguagem Natural, especialmente quando não demandam grande complexidade [Patel and Arasanipalai 2021], optou-se por utilizar esse tipo de arquitetura específica para esta tarefa.

A estrutura deste modelo é apresentada na Figura 7 e inclui as seguintes camadas:

- **Camada de Entrada:** recebe sequências numéricas de comprimento 512, onde os números correspondem à palavras.
- **Camada de *Embedding*:** responsável por transformar a entrada em vetores densos, permitindo a representação semântica das palavras.
- **Camada de *Spatial Dropout*:** utilizada para regularização, descarta aleatoriamente valores durante o treinamento para melhorar a generalização do modelo.
- **Camada Convolucional:** extrai características por meio da aplicação de filtros, identificando padrões em diferentes regiões do *input*.
- **Camada *Global Max Pooling*:** realizando uma operação de *pooling* global máximo, esta camada reduz a dimensionalidade do tensor resultante da camada anterior, extraindo os valores máximos.
- **Camadas Densas:** a primeira camada densa é composta por 155 neurônios, empregando a função de ativação *ReLU*. Já a segunda camada densa possui 1 neurônio e função de ativação *sigmoid*, permitindo a classificação binária.

4.3.3. Classificador de Itens e *Modus Operandi*

A arquitetura selecionada para abordar a tarefa em questão foi a de *BiLSTMs* (*Bidirectional Long Short-Term Memory*). As redes recorrentes bidirecionais oferecem uma vantagem significativa na captura de dependências em dados sequenciais, como no processamento de texto, embora sejam computacionalmente mais custosas. Sua distinção reside na capacidade de considerar informações contextuais tanto do passado quanto do futuro de uma palavra específica em uma sequência. Enquanto as redes recorrentes padrão se concentram apenas nas dependências passadas, as bidirecionais permitem que o modelo capture contextos anteriores e posteriores de maneira abrangente [Goodfellow et al. 2016].

A representação da arquitetura deste modelo está na Figura 8 e é composta pelas seguintes camadas:

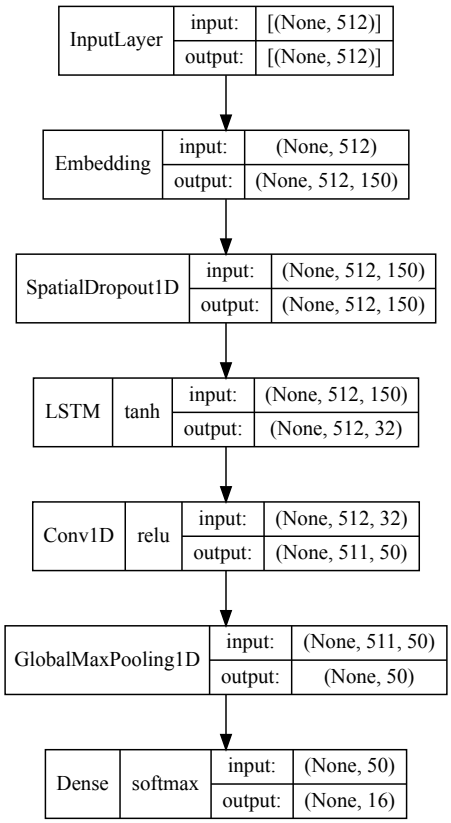

Figura 7. Arquitetura da rede neural para a classificação referente ao tipo de vítima.

- **Camada de Entrada:** recebe sequências numéricas de comprimento 512, onde os números correspondem à palavras.
- **Camada de *Embedding*:** responsável por transformar a entrada em vetores densos, permitindo a representação semântica das palavras.
- **Camada de *Spatial Dropout*:** aplica *dropout* espacial para prevenção de *overfitting*, descartando aleatoriamente valores durante o treinamento para melhorar a generalização do modelo.
- **Camadas *BiLSTMs*:** processam a sequência nos dois sentidos, capturando informações contextuais tanto da esquerda quanto da direita.
- **Camada Densa:** totalmente conectada, produz a saída final do modelo com 107 unidades, correspondendo às classes preditas.

O classificador multirrótulo também enfrenta desafios relacionados ao desbalanceamento extremo entre as classes. Diante desse contexto, a estratégia escolhida para esta tarefa de classificação em questão é a aplicação da função de perda conhecida como *Focal Loss*, mais precisamente, a *Binary Focal Loss*.

A função *Focal Loss* é uma função de perda desenvolvida para lidar com o problema de desequilíbrio de classes em tarefas de detecção de objetos. A característica distintiva da *Focal Loss* é a capacidade de reduzir o impacto das classes majoritárias durante o treinamento, permitindo que o modelo se concentre mais nas classes minoritárias [Lin et al. 2017].

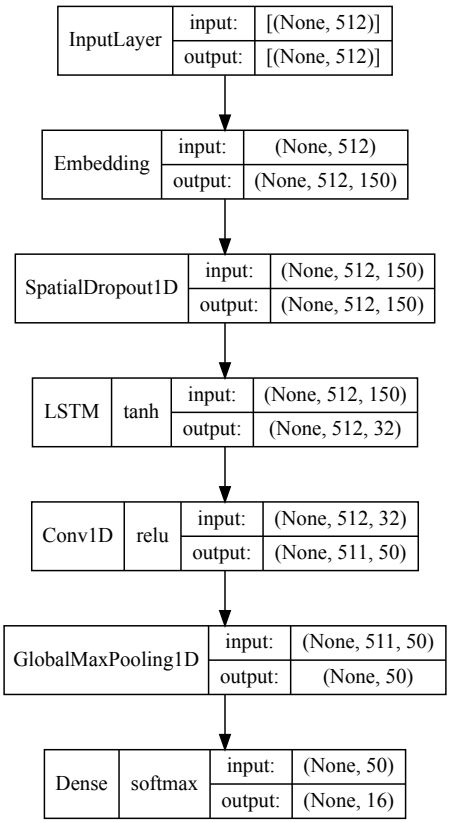

Figura 8. Arquitetura da rede neural para a classificação de *modus operandi* e objetos subtraídos.

5. Resultados

Nessa seção, os resultados da classificação referente a localidade e das classificações referentes a itens, *modus operandi* serão apresentadas primeiramente de forma separada e posteriormente, será feita avaliação da performance em conjunto. Os dados de testes coletados tem aproximadamente 75 mil registros.

5.1. Classificador 1 - Classificação de Localidade

Nesse problema de classificação, acurácia geral obtida foi de 75,86%. No entanto, considerando que a acurácia não é a medida mais adequada para representar a performance para cada classe em casos onde há extremo desbalanceamento entre as classes, também foi calculada a acurácia balanceada que é igual a 73,19%. Para ilustrar a performance entre as diferentes classes, a Figura 9 representa a matriz de confusão para 16 classes do problema, normalizada por linha.

A Tabela 1 apresenta as métricas de acurácia, precisão, *recall* e *F1-score* para todas as classes que representam os locais.

Ao examinar os resultados fornecidos na matriz de confusão, apresentada na Figura 9, e na Tabela 1, observa-se um viés na classificação, no qual certas classes são categorizadas como “sem informação”, destacando-se a classe de crime virtual. Um dos motivos disso pode estar relacionado à natureza intrínseca do crime virtual, que não possui uma localização física específica associada.

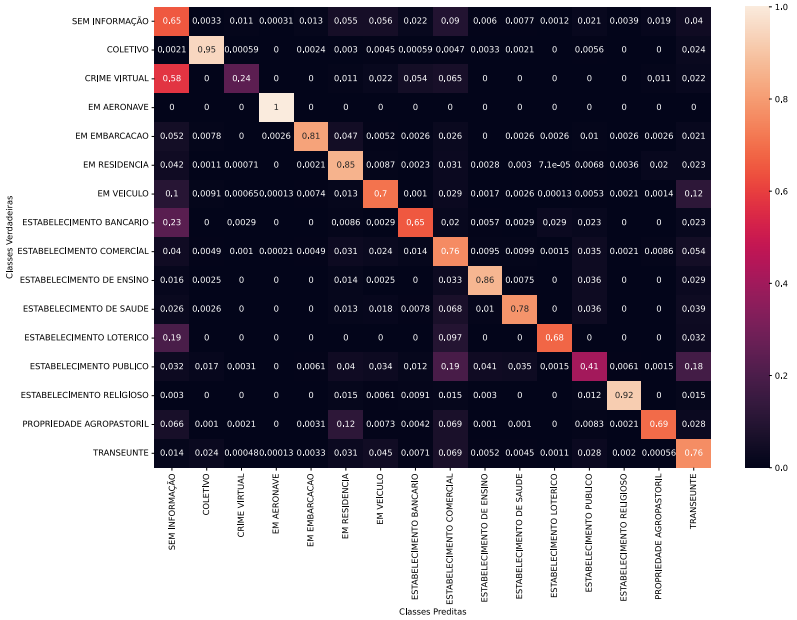

Figura 9. Matriz de confusão do classificador 1.

	ACURÁCIA	PRECISÃO	REVOCAÇÃO	F1-SCORE
COLETIVO	98,76%	81,03%	94,69%	87,33%
EM RESIDÊNCIA	94,52%	85,55%	85,27%	85,41%
TRANSEUNTE	89,30%	87,30%	76,36%	81,47%
ESTABELECIMENTO RELIGIOSO	99,71%	61,26%	92,07%	73,57%
ESTABELECIMENTO DE ENSINO	99,33%	63,91%	85,98%	73,32%
SEM INFORMAÇÃO	90,69%	78,16%	64,96%	70,95%
EM VEÍCULO	94,03%	71,09%	70,42%	70,75%
ESTABELECIMENTO COMERCIAL	91,96%	66,27%	75,86%	70,74%
PROPRIEDADE AGROPASTORIL	98,73%	50,38%	68,85%	58,19%
EM EMBARCAÇÃO	99,37%	43,94%	81,46%	57,09%
ESTABELECIMENTO DE SAÚDE	99,35%	42,86%	77,92%	55,30%
ESTABELECIMENTO BANCÁRIO	98,97%	25,94%	65,14%	37,10%
ESTABELECIMENTO LOTÉRICO	99,89%	23,33%	67,74%	34,71%
ESTABELECIMENTO PÚBLICO	97,48%	15,05%	40,52%	21,95%
EM AERONAVE	99,99%	8,33%	100,00%	15,38%
CRIME VIRTUAL	99,65%	10,23%	23,91%	14,33%

Tabela 1. Métricas para as classes do primeiro classificador.

5.2. Classificador 2 - Classificação de Vítima

O segundo classificador é aplicado exclusivamente a uma porção dos dados de teste, especificamente àquela em que as classes relacionadas à localização correspondem a estabelecimentos ou meios de transporte coletivos, isto corresponde a aproximadamente 15 mil registros. Nessas circunstâncias, a classificação como "em estabelecimento ou em coletivo" indica que uma pessoa teve seus objetos subtraídos nesses locais, enquanto a classificação como "ao estabelecimento ou ao coletivo" significa que a propriedade do estabelecimento ou do coletivo foi o alvo do crime.

A Figura 10 apresenta a matriz de confusão normalizada por linha referente às duas classes neste problema de classificação. E a Tabela 2 apresenta as métricas para cada classe.

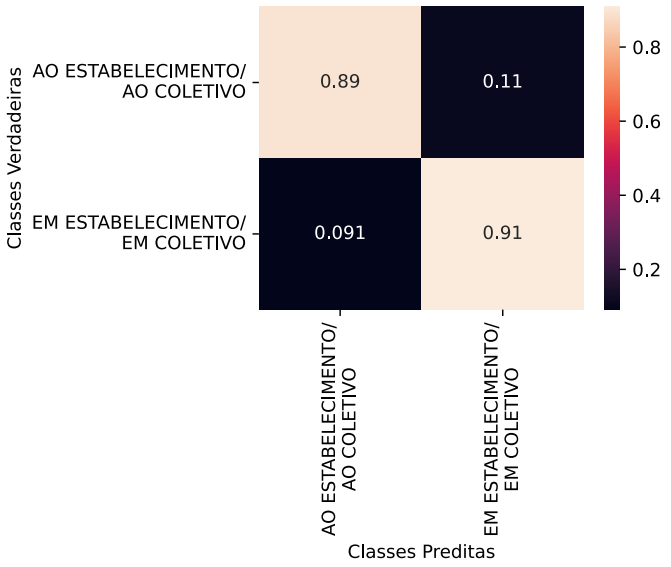

Figura 10. Matriz de confusão do classificador 2.

	ACURÁCIA	PRECISÃO	REVOCAÇÃO	F1-SCORE
AO ESTABELECIMENTO/ AO COLETIVO	90,33%	82,97%	89,24%	85,99%
EM ESTABELECIMENTO/ EM COLETIVO	90,33%	94,43%	90,87%	92,62%

Tabela 2. Métricas para as classes do segundo classificador.

5.3. Classificador 3 - Classificação de Itens e *Modus Operandi*

O terceiro classificador abrange 107 classes. Dada a quantidade, as métricas serão apresentadas apenas para algumas das classes mais relevantes. A Tabela 3 apresenta as métricas destas classes.

O classificador multi-rótulo apresentou 80,56% de acurácia geral e 0,22% de *hamming loss*. Na análise em questão, a acurácia é o indicador que representa a porcentagem de instâncias em que as classes preditas para elas correspondem totalmente às classes verdadeiras.

É importante ressaltar que este é um problema de classificação altamente desbalanceado, onde 53 classes possuem uma quantidade extremamente limitada de registros, muitas delas com menos de uma dezena nos dados de treino. Embora tenham sido aplicadas técnicas para enfrentar o desafio do desbalanceamento, o desempenho do modelo nessas classes com poucos registros é bastante limitado.

Outro ponto relevante a considerar é a variação nas metodologias de classificação ao longo dos quatro anos de registros que compõem a base de treino. Diferenças temporais podem influenciar a eficácia do modelo em diferentes períodos.

Além disso, dado que se trata de uma classificação abrangendo diversas classes para um único registro, é comum que, ao interpretar a descrição textual do incidente, o leitor possa inadvertidamente deixar passar algum detalhe. Portanto, a presença de dados parcialmente incorretos, especialmente no que diz respeito a falsos negativos, é

	ACURÁCIA	PRECISÃO	REVOCAÇÃO	F1-SCORE
CELULAR	96,72%	94,13%	97,83%	95,95%
HIDRÔMETRO	99,81%	94,14%	97,50%	95,79%
BICICLETA	99,31%	86,49%	94,46%	90,29%
MOTO	98,76%	85,27%	93,77%	89,32%
ENERGIA ELÉTRICA	99,77%	83,40%	91,92%	87,45%
TRANSFORMADOR	99,86%	83,78%	90,10%	86,83%
MEDIDOR DE ENERGIA	99,63%	80,92%	89,30%	84,90%
ARMA DE FOGO	99,78%	73,01%	82,56%	77,49%
MEDIANTE FRAUDE	98,89%	65,22%	93,37%	76,79%
APLICATIVO DE TRANSPORTE	99,36%	68,31%	79,38%	73,43%
MOTORISTA	99,53%	63,67%	74,66%	68,73%
COM ARROMBAMENTO	95,05%	58,04%	82,64%	68,19%
CABO	98,64%	52,17%	87,52%	65,37%
VEÍCULO LEVE	99,03%	55,77%	70,31%	62,20%
BOVINO	99,79%	45,93%	93,94%	61,69%
PASSAGEIRO	99,59%	52,55%	69,00%	59,66%
EMBARCAÇÃO	99,70%	46,87%	78,11%	58,58%
CARGA	99,66%	41,94%	75,88%	54,03%
TABAGISMO	99,95%	38,46%	60,00%	46,88%
PIX	99,40%	30,16%	80,18%	43,83%
TÁXI	99,95%	31,82%	60,87%	41,79%
RESTRIÇÃO DE LIBERDADE	99,39%	48,61%	35,12%	40,78%
ALIMENTOS	99,87%	28,13%	51,92%	36,49%
CARTÃO CLONADO	99,95%	36,67%	35,48%	36,07%
VEÍCULO PESADO	99,86%	34,52%	36,71%	35,58%
SAIDINHA BANCÁRIA	99,77%	23,86%	50,60%	32,43%
TENTATIVA DE LATROCÍNIO	99,89%	17,02%	66,67%	27,12%
COMBUSTÍVEL	99,93%	19,57%	37,50%	25,71%
SUÍNO	99,94%	12,24%	60,00%	20,34%
EQUINO	99,98%	20,00%	18,18%	19,05%
APLICATIVO DE ENTREGA	99,90%	26,92%	11,29%	15,91%
PIRATARIA FLUVIAL	99,83%	8,96%	60,00%	15,58%
VIOLÊNCIA DOMÉSTICA	99,72%	18,56%	12,41%	14,88%
GASOLINA	99,95%	11,54%	20,00%	14,63%
COM REFÉM	99,79%	11,21%	16,44%	13,33%
COM LESÃO	99,86%	14,04%	12,50%	13,22%
MOTOTÁXI	99,86%	9,09%	17,07%	11,86%
ANIMAL DE ESTIMAÇÃO CACHORRO	99,98%	50,00%	5,56%	10,00%
OUTROS	96,71%	35,53%	1,10%	2,14%
JOIAS	99,98%	0,00%	0,00%	0,00%

Tabela 3. Métricas para as classes do terceiro classificador.

significativa tanto nos dados de treino quanto nos dados de teste.

5.4. Performance Conjunta

Em uma avaliação conjunta, os resultados dos classificadores são tratados como se fossem um sistema único. Nesse contexto, a acurácia representa a porcentagem de instâncias cujas classes preditas são exatamente iguais às classes verdadeiras, refletindo a correspondência exata entre as predições e os resultados reais. Em conjunto, os múltiplos classificadores alcançam uma acurácia geral de 62,78%. Entretanto, essa métrica não fornece uma visão sobre a quantidade de valores que podem estar parcialmente corretos.

Dado o contexto desse cenário, no qual múltiplas classes estão associadas a um único registro, existe uma considerável probabilidade de classificação parcialmente correta. Com esse entendimento, também foi calculado o índice de similaridade de Jaccard para cada registro, sendo que a média desse índice em toda a base de teste foi de 73,45%. Este índice oferece uma medida mais abrangente, levando em consideração a sobreposição de elementos entre os conjuntos de valores verdadeiros e preditos.

6. Conclusão

A categorização de dados na esfera da segurança pública vai muito além de simples objetivos estatísticos e divulgação. Seu papel essencial se evidencia ao proporcionar uma contribuição significativa para a análise, monitoramento e gestão no contexto da segurança. Esta pesquisa buscou uma abordagem para a classificação automática de dados textuais referentes a crimes de furtos e roubos. E apresentou resultados notáveis, obtendo um desempenho satisfatório nos dados de teste, com uma acurácia geral de 62,78% e um coeficiente de Jaccard igual a 73,45%. Embora esses resultados sejam bons, considerando a complexidade da tarefa, ainda há margem para melhorias, especialmente no que diz respeito ao desbalanceamento extremo dos dados, onde algumas classes contam com menos de uma dezena de registros na base.

É crucial reconhecer a limitação dos classificadores em um ambiente de produção e evitar depender exclusivamente das respostas dos classificadores para classes com precisão ou revocação extremamente baixas.

Assim, é imprescindível adotar uma abordagem cautelosa e complementar a classificação da ferramenta com outras formas de validação e revisão. Uma possível estratégia para lidar com classes com baixa revocação é realizar uma busca automática por termos associados a essas classes. Já quando há uma classificação positiva em uma classe com baixa precisão, uma forma de validação é a revisão humana para confirmar o resultado. Considerando que o desempenho dos classificadores em cada classe está diretamente relacionado à quantidade dessas classes nos dados de treinamento, a quantidade de relatos a serem selecionados para leitura tende a ser limitada.

Por outro lado, em classes onde ambas as métricas são altas, as ferramentas podem fornecer uma avaliação confiável do conteúdo dos relatórios, oferecendo uma estimativa precisa da prevalência dessas classes na base de relatos.

Além disso, é viável desenvolver estratégias para lidar com mudanças dinâmicas na classificação dentro do órgão, visando padronizar a base de dados. Aprimorar técnicas para lidar com erros de classificação, especialmente falsos negativos em tarefas de classificação multirrótulo, é crucial para pesquisas futuras.

Outra possível direção para futuras pesquisas é a exploração do uso de ferramentas de otimização de hiperparâmetros, como o Keras Tuner, onde se pode aplicar técnicas como busca aleatória, busca em grade ou otimização bayesiana, para descobrir combinações ideais de hiperparâmetros que otimizem a acurácia e a generalização dos classificadores.

Além disso, uma área de investigação possível para estudos futuros é a utilização de modelos pré-treinados baseados em *Transformers* . Seria relevante avaliar o ganho de desempenho desses modelos específicos neste domínio.

Agradecimentos

A primeira autora deste trabalho expressa sua gratidão ao Prof. Dr. Filipe Saraiva pela orientação e contribuições que foram fundamentais tanto para o desenvolvimento e conclusão deste estudo quanto para o curso de graduação como um todo. Além disso, gostaria de agradecer aos professores do curso de Ciência da Computação da Universidade Federal do Pará, cujos ensinamentos foram fundamentais para o desenvolvimento deste trabalho. Por fim, também expressamos nossos agradecimentos aos profissionais da SIAC, cuja cooperação foram inestimáveis para a realização desta pesquisa.

Referências

Abadi, M. et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Abubakar, H. D., Umar, M., and Bakale, M. A. (2022). Sentiment classification: Review of text vectorization methods: Bag of words, tf-idf, word2vec and doc2vec. *SLU Journal of Science and Technology*, 4(1 & 2):27–33.

Birks, D., Coleman, A., and Jackson, D. (2020). Unsupervised identification of crime problems from police free-text data.

Caseli, H. M. and Nunes, M. G. V., editors (2023). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN. `https://brasileiraspln.com/livro-pln`.

Chollet, F. et al. (2015). Keras.

Das, P., Das, A. K., Nayak, J., Pelusi, D., and Ding, W. (2019). A graph based clustering approach for relation extraction from crime data. *IEEE Access*, 7:101269–101282.

dos Reis Matos, H. M. (2022). *Um Classificador Supervisionado para Relatos Policiais no Estado do Pará*. Belém.

Fórum Brasileiro de Segurança Pública (2023). *17º Anuário Brasileiro de Segurança Pública*. Fórum Brasileiro de Segurança Pública. `https://forumseguranca.org.br/wp-content/uploads/2023/07/anuario-2023.pdf`.

Goodfellow, I. J., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press, Cambridge, MA, USA.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spacy: Industrial-strength natural language processing in python.

Johnson, J. M. and Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):27.

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Kuang, D., Brantingham, P., and Bertozzi, A. (2017). Crime topic modeling. *Crime Science*, 6.

Leskovec, J., Rajaraman, A., and Ullman, J. D. (2014). *Mining of Massive Datasets*. Cambridge University Press, USA, 2nd edition.

Li, Q., Zhang, Q., Yao, J., and Zhang, Y. (2020). Event extraction for criminal legal text. In *2020 IEEE International Conference on Knowledge Graph (ICKG)*, pages 573–580.

Li, X. and Ning, H. (2020). Chinese text classification based on hybrid model of cnn and lstm. In *Proceedings of the 3rd International Conference on Data Science and Information Technology*, DSIT 2020, page 129–134, New York, NY, USA. Association for Computing Machinery.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollar, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Luan, Y. and Lin, S. (2019). Research on text classification based on cnn and lstm. In *2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*. IEEE.

Maimon, O. and Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook, 2nd ed.*

Oliveira, B. S. N. (2020). *Aprendizado Profundo para Reconhecimento de Entidades Nomeadas em Narrativas de Roubos*. Quixadá.

Patel, A. and Arasanipalai, A. (2021). *Applied Natural Language Processing in the Enterprise*. O'Reilly Media.

Qi, Z. (2020). The text classification of theft crime based on tf-idf and xgboost model. In *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*.

SEGUP-PA (2023). Institucional - Portal da Transparência da Segurança Pública. `http://sistemas.segup.pa.gov.br/transparencia/institucional/`. Acessado em 18 de Dezembro de 2023.

Wang, K., Zhang, P., and Su, J. (2020). A text classification method based on the merge-lstm-cnn model. *Journal of Physics: Conference Series*, 1646(1):012110.