Revista Brasileira de Computação Aplicada, Abril, 2026

DOI: 10.5335/rbca.v18i1.17029
Vol. 18, Nº 1, pp. 68–80
Homepage: seer.upf.br/index.php/rbca/index

ARTIGO ORIGINAL

Avaliação de Desempenho de Modelos BERT para Classificação Automatizada de Boletins de Ocorrência na Cidade de Marabá-PA

Performance Evaluation of BERT Models for Automated Police Report Classification in Marabá-PA

Marcílio D. S. Marques [1], Reginaldo C. dos Santos Filho [2], Hugo P. Kuribayashi [3], Adam D. F. dos Santos [4], Anderson da Silva Soares [5]

[1,3,4]University of Southern and Southeastern Para, [2]Federal University of Para, [5]Federal University of Goias

*marciliodsm@unifesspa.edu.br; regicsf@ufpa.br; hugo@unifesspa.edu.br; adamdreyton@unifesspa.edu.br; andersonsoares@ufg.br

Recebido: 09/05/2025. **Revisado:** 30/03/2026. **Aceito:** 26/04/2026.

Resumo

Este trabalho aborda o desenvolvimento de um classificador para boletins de ocorrência da cidade de Marabá-PA. Utilizaram-se técnicas de mineração de dados e ajuste fino de Linguagem de Language Models (LLMs) pré-treinados, como o Bidirectional Encoder Representations from Transformers (BERT) e sua versão adaptada ao português, o BERTimbau. A avaliação dos modelos indica que os transformadores BERT base e BERTimbau alcançaram acurácias globais de aproximadamente 90% e 92%, respectivamente, em experimentos realizados com dados de teste. Esses resultados demonstram a viabilidade do uso de LLMs para a classificação automática de boletins de ocorrência, oferecendo potencial para aprimorar a análise de dados criminais e contribuir para políticas de segurança pública mais eficientes e orientadas por dados.

Palavras-Chave: Aplicações de IA; Aprendizado de Máquina; Boletins de Ocorrência; Classificação; Violência.

Abstract

This work addresses the development of a classifier for police reports from the city of Marabá-PA, employing data mining techniques and fine-tuning of pre-trained Large Language Models (LLMs), such as Bidirectional Encoder Representations from Transformers (BERT) and its Portuguese-adapted version, BERTimbau. The evaluation of the models indicates that the BERT base and BERTimbau transformers achieved overall accuracies of approximately 90% and 92%, respectively, in experiments conducted with test data. These results demonstrate the feasibility of using LLMs for the automatic classification of police reports, offering potential to enhance criminal data analysis and contribute to more efficient, data-driven public security policies.

Keywords: Applications of AI; Classification; Machine Learning; Police Reports; Violence.

1 Introdução

A segurança pública no Brasil enfrenta desafios persistentes, refletidos nos elevados índices de violência e criminalidade. Entre os fenômenos mais alarmantes está o aumento das Mortes Violentas Intencionais (MVI), que são um reflexo da complexidade dos problemas sociais e estruturais enfrentados pelo país (Fórum Brasileiro de Segurança Pública, 2024). Este tipo de crime resulta em tragédias pessoais e desencadeia um impacto profundo nas comunidades,

perpetuando um ciclo de medo e instabilidade (Eufrazio, 2024).

As MVIs impactam profundamente a dinâmica social e econômica das comunidades afetadas, além de criarem um ambiente de incerteza que prejudica a coesão social e a confiança nas instituições (Carneiro, 2022). A presença constante de violência desestabiliza a vida cotidiana dos cidadãos e inibe investimentos e o oferecimento de serviços essenciais. Assim, o desenvolvimento local é comprometido, potencialmente perpetuando condições que podem estar associadas às causas das MVIs.

Dados do 18º Anuário Brasileiro de Segurança Pública 2024 indicam que entre os anos de 2019 e 2023 foram registradas 240.790 MVIs, representando uma média anual de 48.158 MVIs por ano (Fórum Brasileiro de Segurança Pública, 2024), conforme demonstrado pela Tabela 1. Neste contexto, é importante verificar que estes patamares de MVIs registrados em contextos urbanos no Brasil se assemelham às fatalidades contabilizadas em cenários de guerra. Este panorama evidencia a gravidade da crise de segurança pública que afeta certas regiões no país, colocando em xeque a eficácia das políticas de intervenção e proteção social.

Tabela 1: Série histórica de MVIs no Brasil.

Ano	2019	2020	2021	2022	2023	Total	Média
MVI	47.765	50.448	48.286	47.963	46.328	240.790	48.158

Em resposta a essa realidade alarmante, é imperativo que políticas públicas sejam direcionadas para a promoção da paz e segurança. Em observância ao Objetivo de Desenvolvimento Sustentável (ODS) n. 16, que visa "promover sociedades pacíficas e inclusivas para o desenvolvimento sustentável, proporcionar acesso à justiça para todos e construir instituições efetivas, responsáveis e inclusivas em todos os níveis", a implementação de estratégias eficazes para a redução das taxas de criminalidade e promoção da segurança pública é essencial.

No estado do Pará, a situação não é diferente. O município de Marabá, com uma população aproximada de 221.050 habitantes e com o quarto maior PIB per capita do Estado (IBGE, 2022; FAPESPA, 2024), registrou um aumento de 10,11% nos boletins de ocorrência entre 2022 e 2023, destacando a necessidade premente de estratégias eficazes para o enfrentamento da criminalidade no município (Fórum Brasileiro de Segurança Pública, 2024).

A dinâmica socioeconômica de Marabá, impulsionada pelo crescimento industrial e migração, contribui para desafios urbanos e sociais que influenciam os índices criminais. Órgãos como a Secretaria Adjunta de Inteligência e Análise Criminal (SIAC), responsável pelo processamento de dados de segurança no Pará, enfrentam dificuldades na gestão e análise desses registros devido à limitação de recursos e dependência de softwares licenciados que não atendem plenamente às necessidades locais. Este cenário compromete a identificação de padrões de criminalidade e a formulação de respostas rápidas e eficazes, intensificando assim os prejuízos à população local.

Além disso, o município representa um centro econômico e logístico na Amazônia Sul-Oriental, embora enfrente desafios significativos relacionados à criminalidade, incluindo roubo, tráfico de drogas e outros crimes violentos, conforme relacionado pela Fig. 1, que detalha a série histórica de registro de Boletins de Ocorrência Policial (BOP) no município. O recorte destaca os 5 (cinco) tipos de crimes registrados com maior prevalência, sendo possível verificar que, além dos delitos considerados violentos, existe uma grande predominância dos ilícitos relacionados à lesão ao patrimônio dos indivíduos.

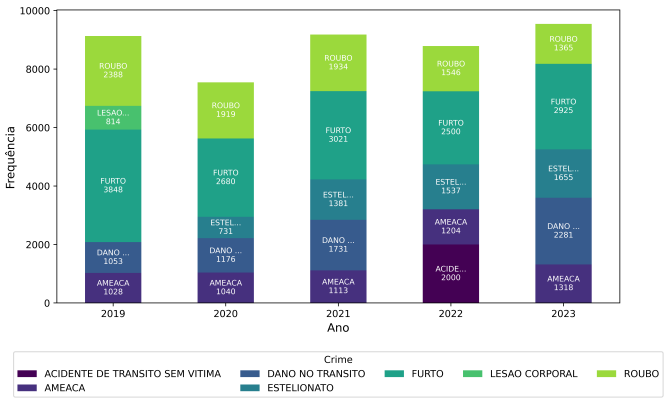

Figura 1: Série histórica de registros de BOPs por ano em Marabá.

No que se refere ao volume de dados de BOPs registrados na cidade de Marabá, a SIAC encontra-se incapaz de consolidar todos os tipos de crimes, optando-se apenas por algumas categorias para a classificação, uma vez que tradicionalmente os BOPs abertos pelos cidadãos são classificados manualmente por humanos. Outro desafio dessa secretaria é a utilização de softwares licenciados, os quais, em sua maioria, não satisfazem plenamente suas exigências (Souza, 2022). A ausência de métodos tecnológicos eficazes impacta negativamente a organização, especialmente em setores ligados à segurança pública. Para mitigar esses desafios, é necessário adotar ferramentas tecnológicas apropriadas, possibilitando tanto a alocação estratégica de profissionais nos setores que demandam maior especialização quanto a mobilização de viaturas para as áreas mais necessitadas.

Por outro lado, a classificação de dados em registros criminais enfrenta desafios significativos, especialmente devido à variabilidade na redação dos relatos, à ambiguidade de termos e à necessidade de contextualização para uma interpretação precisa. Estes relatos são considerados fontes de dados não estruturados (informações não organizadas em um modelo de dados pré-definido) e redigidos em livre formato, que potencialmente apresentam inconsistências. Métodos convencionais frequentemente apresentam limitações ao lidar com essas complexidades, tornando essencial a aplicação de abordagens mais avançadas. Nesse sentido, técnicas de Aprendizado de Máquina (AM) emergem como uma solução promissora para aprimorar a análise desses dados, permitindo maior precisão e adaptabilidade na identificação de padrões e tendências (Devlin et al., 2018).

Nesse contexto, a aplicação de métodos de AM e de Grandes Modelos de Linguagem - *Large Language Model* (LLM) surge como alternativa para otimizar a análise de BOPs, aprimorando a classificação de crimes e a detecção de tendências criminais. A literatura relacionada indica que modelos como o *Bidirectional Encoder Representations from Transformers* (BERT) já demonstraram eficiência na extração e compreensão de dados textuais complexos, tornando-se uma abordagem promissora para fortalecer a infraestrutura de segurança pública e a gestão de informação em Marabá. Assim, este trabalho busca realizar a avaliação de modelos BERT, aplicados para análise e classificação de BOPs em Marabá-PA.

Com isso, o objetivo é viabilizar uma análise baseada nos dados, isenta de viés, que preserve a qualidade do julgamento humano enquanto aproveita a rapidez dos processos automatizados, desde que seja assegurada a sua confiabilidade. Além disso, almeja-se que essa abordagem possa complementar as metodologias tradicionais de tratamento dos dados de segurança pública, auxiliando em tarefas estatísticas, reduzindo esforços humanos manuais e repetitivos, e contribuindo para aprimorar a qualidade das informações formalizadas nos sistemas de bancos de dados públicos do estado do Pará. As principais contribuições dessa pesquisa são:

i. Investigação do desempenho dos modelos BERT e BERTimbau na classificação de BOPs, em um contexto de múltiplas classes de crimes, explorando um aspecto que ainda carece de exploração na literatura;
ii. Apresentação de uma metodologia que incorpora técnicas avançadas de processamento de linguagem natural - *Natural Language Processing* (NLP) para o pré-processamento das descrições das ocorrências, visando extrair informações relevantes e melhorar a qualidade do treinamento dos modelos;
iii. Utilização de testes estatísticos como o de Friedman e os post-hoc de Shaffer e Bergmann para validar diferenças de desempenho entre os modelos, fornecendo uma avaliação mais robusta do impacto dos modelos na classificação de crimes;

O restante deste artigo está organizado da seguinte forma: a Seção 2 relaciona trabalhos semelhantes que usam ferramentas NLP para resolução de variados problemas similares ao estabelecido neste trabalho. A Seção 3 apresenta a metodologia adotada, incluindo a coleta de dados, as fases de desenvolvimento do modelo, a modelagem baseada no BERT e os critérios de avaliação. Na Seção 4, são discutidos os resultados obtidos, abordando métricas de desempenho, análise da matriz de confusão e testes estatísticos não-paramétricos. Por fim, a Seção 5 traz as conclusões do estudo.

2 Trabalhos Correlatos

Diversos trabalhos na literatura correlata têm explorado a aplicação de técnicas de NLP em uma variedade de domínios, incluindo análise de sentimentos, identificação de conteúdos ofensivos e sumarização de textos. Essas pesquisas utilizam diferentes modelos, como BERT e suas variantes, aplicando-os a bases de dados específicas para extrair informações pertinentes e melhorar a compreensão de dados não-estruturados. Além disso, os estudos investigam a eficácia de metodologias distintas de pré-processamento e avaliação estatística, demonstrando desafios na categorização e análise de grandes volumes de texto. Essas abordagens fornecem uma base para o desenvolvimento de novas soluções, evidenciando a relevância e a aplicabilidade das técnicas de NLP em contextos variados.

No trabalho de Oliveira (2023), o autor investiga as emoções expressas em redes sociais durante eventos relevantes, utilizando *word embeddings*, redes neurais recorrentes e *transformers*. Os resultados indicaram uma correlação entre as datas dos eventos e as variações emocionais, embora a análise de sentimentos apresente desafios devido a fatores contextuais e vieses nos dados. Destaca-se a necessidade de aprimoramento contínuo para uma compreensão mais precisa do comportamento humano nas redes. Diversos modelos foram avaliados para a classificação de emoções, incluindo *Bi-Directional Gated Recurrent Unit* (bi-GRU), *FastText*, BERT e *DistilBERT*, sendo o BERT identificado como o mais promissor, com melhor desempenho médio na métrica F1-Score.

No trabalho de Barros (2022), avaliam-se as diversas abordagens para a sumarização extrativa de documentos textuais, baseadas no modelo *BERTSUM*, utilizando uma abordagem multientrada para lidar com documentos de diferentes tamanhos e domínios. Os autores também avaliam o uso do modelo BERT para sumarização extrativa, propondo abordagens que superam essa limitação e podem lidar com documentos de comprimento variável. Além disso, uma técnica de agrupamento automático foi sugerida para tratar documentos de sub-domínios distintos. Os resultados demonstram a eficácia dessas abordagens, adaptando-se às características dos conjuntos de dados.

No trabalho de Bomfim and Lopes (2022), os autores avaliam o desempenho de algoritmos de mineração de texto treinados para identificar mensagens ofensivas relacionadas a *ciberbullying*. O trabalho considera o uso dos modelos BERT e *One Class SVM*, treinados com duas bases de dados contendo "tweets" em língua portuguesa. O modelo BERT, utilizando o *BERTimbau* pré-treinado para língua portuguesa, alcançou um F1-Score total de 80% na primeira base, com precisão e recall também de 80%. O *One Class SVM* obteve um F1-Score global de 61%, com precisão de 67%. Na segunda base de dados, o BERT atingiu um F1-Score de 67%, enquanto o outro modelo obteve 48%. Além disso, ficou evidente a dificuldade dos modelos em identificar textos com sentido figurado, sarcasmo ou ironia.

A pesquisa de Barros et al. (2021) apresenta a aplicação de diferentes modelos de sumarização automática de documentos textuais no domínio das "Notícias Crimes". O trabalho adota o modelo BERT para sumarização extrativa de documentos textuais em português, além de propor a utilização do modelo de BertSumPor, um modelo que consegue lidar com documentos com variações de tamanho e complexidade. Os resultados indicam que o modelo de BertSumPor com entrada variada obteve os melhores resultados da métrica *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE).

Além disso, no trabalho de Roy and Goldwasser (2021),

os autores avaliam a relação entre o uso de fundamentos morais pelos políticos em redes sociais e o posicionamento deles em relação a questões como controle de armas e imigração. Para realizar as análises propostas neste estudo, foi aplicada uma abordagem denominada *Deep Relational Learning* (DRaiL) para a captura de moralidade em texto. O BERT, neste caso, é usado para criar as representações dos *tweets* por meio de *embeddings* do próprio *transformer*.

No trabalho de Yu et al. (2024) propõe-se uma variante modificada do BERT (M-BERT), para a detecção de URLs maliciosos, integrando informações semânticas e atributos específicos das URLs. O M-BERT apresenta nova camada de *embedding* que incorpora características como comprimento, presença de números e propriedades dos domínios, aprimorando a representação semântica. Os experimentos demonstraram que o M-BERT atingiu 94,42% de precisão, superando modelos tradicionais como *Convolutional Neural Network* (CNN) e Long Short Term Memory (LSTM), além de outras variantes de LLMs.

No trabalho de Silveira et al. (2023) os autores apresentam o modelo LegalBert-pt, desenvolvido especificamente para o domínio jurídico em língua portuguesa, utilizando 1,5 milhão de documentos legais brasileiros. Foram criadas duas versões do modelo: uma treinada do zero (LegalBert-pt SC) e outra baseada no *BERTimbau* (LegalBert-pt FP). Avaliações intrínsecas e extrínsecas demonstraram que os modelos especializados superaram significativamente modelos genéricos, como o BERTimbau-Base, em tarefas de Named Entity Recognition (NER) e classificação de texto, destacando a vantagem do pré-treinamento em corpus específico. O LegalBert-pt FP obteve os melhores resultados em termos de perplexidade e F1-Score, reforçando a importância de modelos especializados para tarefas específicas.

No trabalho de Matos et al. (2022), os autores desenvolveram um classificador supervisionado de boletins policiais utilizando informações da SIAC, empregando CNN para processar os relatórios e prever o tipo de evento. Esse modelo foi aplicado para agilizar processos estatísticos e realizar análises qualitativas automatizadas sobre grandes volumes de dados. Diferentemente dessa abordagem baseada em redes convolucionais, este trabalho adota o Modelo BERT para a classificação de crimes, explorando sua capacidade de compreender o contexto linguístico dos boletins de ocorrência e aprimorar a precisão na categorização dos incidentes relatados.

A revisão de literatura evidencia diversas possibilidades de uso das tecnologias de AM e LLM no tratamento de texto. Os estudos revisados nesta seção abordam uma variedade de aplicações de modelos de linguagem, como BERT, em diferentes domínios e tarefas. Entre essas tarefas estão: a análise de emoções em redes sociais; a sumarização extrativa de documentos textuais; a identificação de mensagens ofensivas relacionadas ao *ciberbullying*; a sumarização de "notícias crimes"; entre outros. Diante desse contexto, esta pesquisa busca comparar diferentes variantes do modelo BERT para identificar a mais adequada para a tarefa de classificação de BOPs da cidade de Marabá-PA, contribuindo assim para o avanço do conhecimento nessa área.

3 Materiais e Métodos

A Fig. 2 apresenta uma representação do desenho metodológico adotado neste trabalho destinado à concepção de um classificador supervisionado de BOPs baseado no uso de LLMs. Este classificador busca analisar ocorrências para um determinado conjunto de classes de crimes (roubo, homicídio, etc), e por esse motivo, o escopo deste estudo abrange a implementação e avaliação de técnicas de NLP dedicadas ao reconhecimento e identificação de regras linguísticas presentes nos textos estudados.

No interesse de garantir a confiabilidade no processo, aqui baseado no *Cross Industry Standard Process for Data Mining* (CRISP-DM), as bases de dados passaram pelas fases de Seleção, Limpeza e Transformação para posterior carga e uso no treinamento do modelo LLM. Depois da fase de Transformação os dados foram inseridos na Modelagem, ou seja, o modelo escolhido foi treinado usando os dados advindos da fase anterior.

Na conclusão do processo, foi executada a fase de Avaliação. Neste ponto, o trabalho usou métricas de avaliação de resultados para classificação multi-classe a fim de validar o desempenho do modelo proposto. Entre as métricas estão: acurácia, precisão, *recall* (sensibilidade), F1-score, entre outras. Ainda na fase de Avaliação, foi realizada a análise de desempenho dos resultados entre duas variantes do modelo BERT com o objetivo de avaliar cada uma e, com isso, classificar o melhor *Transformer* mediante um ranking baseado em estudos estatísticos não-paramétricos.

3.1 Coleta de Dados Brutos

No que diz respeito à coleta dos dados empregados na análise, a SIAC disponibilizou um conjunto de registros policiais referentes aos anos de 2019 até 2023. Esses registros são armazenados em arquivos de formato tabular, totalizando cerca de 1,95 gigabytes de dados combinados, os quais consistem em 2.342.039 amostras e 92 atributos. Esses atributos incluem colunas relacionadas à identificação do registro, data e hora, localização, descrição do *modus operandi*, informações sobre vítimas e autores, bem como o relato detalhado do evento. No que concerne às colunas de tipologia criminal, é relevante destacar os seguintes atributos:

- *"relato"*: seção mais relevante do BOP, onde é feita uma narrativa detalhada do incidente, incluindo testemunhos, descrição dos danos ou lesões, e quaisquer outras informações relevantes para a compreensão do ocorrido. É um atributo não estruturado e composto por informações essenciais para fornecer um relato abrangente e detalhado do evento em questão, possibilitando às autoridades competentes investigar e tomar as medidas necessárias conforme o caso reportado;
- *"registros"*: classes de eventos atribuídas nas delegacias no momento do registro do BOP. Devido à natureza emergencial do preenchimento, nem sempre refletem com precisão o evento descrito no relato, e não são normalizados, apresentando casos nos quais as classes remetem ao Código Penal Brasileiro (por exemplo, "ART. 147 - AMEAÇA"), e até mesmo subconjuntos de classes com erros ou duplicidades de grafia que podem ser agru-

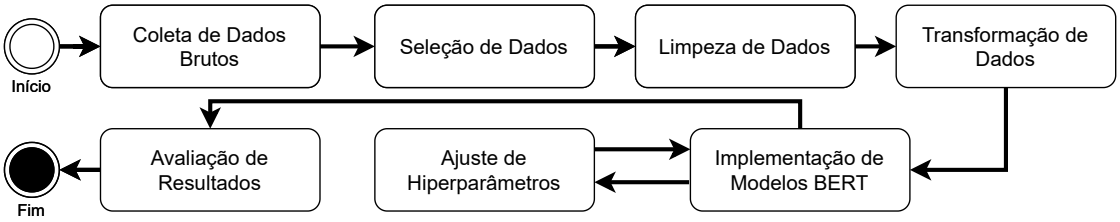

Figura 2: Metodologia de trabalho adotada para classificação de boletins.

pados em uma única classe (por exemplo, "ART. 157 - FURTO", "ARTIGO 157 - FURTO", "FUERTO", "FURTO", etc.);

- *"consolidado"*: classes de eventos atribuídas na base de dados da SIAC, visando uma alternativa mais precisa e confiável em comparação com as classes de registros. Essas classes são principalmente utilizadas na geração de relatórios estatísticos para os gestores da Segurança Pública no Estado do Pará, na mídia interessada em dados quantitativos e nos portais de transparência da secretaria.

Cabe complementar que em relação à tipologia criminal, as colunas acima foram selecionadas como sendo as mais relevantes para a tarefa de classificação proposta no trabalho. Normalmente, nesse tipo de tarefa, são fornecidas duas colunas essenciais: "texto" e "etiqueta", sendo que esta última corresponde à classificação do "texto". Ao comparar essas colunas com as do conjunto de dados de BOPs, optou-se por utilizar os atributos "relato" e "consolidado" como os mais relevantes para o treinamento e teste do modelo de classificação.

Segundo o trabalho de Matos et al. (2022), no contexto operacional interno da SIAC, a coluna de "consolidado" emerge como uma solução para a baixa consistência da coluna de "registros", cujo uso direto poderia comprometer as demandas estatísticas da secretaria. O departamento de estatística dessa secretaria designa um grupo de quinze analistas criminais para ler e rotular os relatos em uma classe específica, com base em seus conhecimentos da legislação brasileira e, em casos complexos, na deliberação coletiva desses analistas, resultando nas classes de "consolidado". Atualmente, essa análise é realizada em um conjunto selecionado de leituras de relatos, concentrando-se principalmente em crimes violentos (homicídio, roubo, estupro, etc.), devido ao grande volume de dados diários gerados nas delegacias. Essa metodologia indica que cerca de 10% das classes de "registros" não são compatíveis com as classes finais de "consolidado", uma margem de erro que é o principal foco de alocação de esforços para a melhoria da qualidade dos dados.

Conforme exposto, dada a carência por uma ferramenta automática de confirmação das classes de eventos criminais para processamento do volume massivo de registros policiais diários, em conjunto com a necessidade de definir um conjunto descritivo das classes de relatos para o problema de pesquisa, o atributo "consolidado" foi escolhido como a coluna-alvo. Essa escolha se justifica também pela confiabilidade desta coluna, uma vez que ela é preenchida pelos analistas depois de fazerem uma análise das colunas "registro" e "relato".

3.2 Seleção de Dados

A partir da fase de coleta de dados brutos, esta fase busca identificar e escolher os registros mais relevantes para o treinamento do modelo. Durante essa etapa, filtros são aplicados nas amostras (linhas) que melhor se alinham aos objetivos da pesquisa, garantindo que o conjunto final de dados seja adequado e representativo para as análises subsequentes.

O primeiro filtro de seleção feito na base de dados foi a localização geográfica dos BOPs que ocorreram especificamente no município de Marabá-PA. Para isso, a metodologia deste trabalho adota o atributo "municípios" presente nas tabelas, para selecionar somente as amostras desejadas.

O filtro selecionou as amostras pertencentes aos 10 (dez) maiores grupos de crimes registrados na referida cidade. Para atingir esse objetivo, inicialmente, o *dataset* foi agregado segundo o atributo "consolidado", que classifica os eventos em distintas categorias, como furto, roubo e dano. Posteriormente, os grupos resultantes foram organizados em ordem decrescente com base na quantidade de amostras. Dessa forma, tornou-se viável selecionar as amostras pertencentes aos dez grupos de crimes mais representativos em Marabá, conforme ilustrado na Tabela 2. A justificativa para a aplicação deste filtro final reside na necessidade de assegurar a manutenção de um número mínimo de registros para cada classe de crime. Uma quantidade excessivamente reduzida de amostras poderia comprometer a adequação do treinamento do modelo. Adicionalmente, é importante destacar que as classes de crimes listadas na Tabela 2 estão em conformidade com o Código Penal (CP) (Brasil, 1940) e o Código de Trânsito Brasileiro (CTB) (Brasil, 1997).

No fim desta fase, apenas 2 (duas) colunas foram julgadas como relevantes para o desenvolvimento do modelo proposto: "relato" e "consolidado". Estas colunas foram selecionadas devido ao seu potencial ganho de informação e capacidade de fornecer informações cruciais para a identificação e classificação precisa dos BOPs, garantindo assim a maior eficácia dos modelos avaliados.

3.3 Limpeza de Dados

Esta fase é fundamental para elevar a qualidade do conjunto de dados, tornando possível a aplicação de técnicas de mineração de dados de forma eficaz e produtiva. Ao

Tabela 2: Classes de crimes mais comuns em Marabá entre 2019 e 2023.

Classes	Registros
FURTO (CP, Art 155)	14.974
ROUBO (CP, Art 157)	9.152
DANO NO TRÂNSITO (CTB, Art 293)	6.289
AMEAÇA (CP, Art 147)	5.703
ESTELIONATO (CP, Art 171)	5.605
LESÃO CORPORAL (CP, Art 129)	3.870
ACIDENTE DE TRÂNSITO SEM VÍTIMA (CTB, Art 305)	2.670
LESÃO NO TRÂNSITO (CTB, Art 303)	1.437
INVASÃO DE DISPOSITIVO INFORMÁTICO (CP, Art 154-A)	1.162
DANO (CP, Art 163)	974
TOTAL	51.836

analisar a base de dados, é possível identificar a presença de valores duplicados, o que é comum em conjuntos de dados que contêm informações sobre eventos complexos, como ocorrências criminais envolvendo múltiplas vítimas ou autores. Por exemplo, existem algumas situações em que um único evento criminal pode ser registrado em mais de um boletim de ocorrência, resultando na duplicação de informações.

Para lidar com essa questão, a metodologia de limpeza adotada neste trabalho considera o campo "nro_bop" como identificador único dos BOPs. Tal campo nos permite identificar e remover eficientemente as duplicatas, preservando apenas a primeira ocorrência de cada evento. Durante o processo de remoção de duplicatas, além da coluna "nro_bop", considerou-se também o campo "relato", que descreve detalhes do evento. Dessa forma, todas as instâncias duplicadas, tanto na coluna "nro_bop" quanto no campo "relato", foram devidamente eliminadas, mantendo a integridade e a qualidade dos dados.

3.4 Transformação de Dados

A fase de transformação marca a conclusão da preparação dos dados, na qual são gerados atributos derivados, novos registros ou valores transformados para atributos existentes, resultando na forma final dos dados que serão submetidos à etapa de modelagem. Durante essa etapa, diversas alterações são aplicadas sobre a coluna de relatos, visando aprimorar a qualidade e a consistência dos dados. Essas alterações incluem:

- **Transformação dos relatos para caixa baixa:** Todos os caracteres dos relatos são normalizados para minúsculo, garantindo uniformidade na representação dos textos;
- **Remoção de tags HTML:** Os relatos na base de dados originais podem conter elementos de marcação HTML oriundos do sistema que realiza o registro dos BOPs. Esses elementos de marcação são indesejados e podem introduzir ruídos nos dados, sendo identificados e eliminados, preservando apenas o conteúdo textual relevante dos relatos;
- **Remoção de espaços múltiplos:** Qualquer ocorrência de espaços múltiplos nos relatos é removida, assegurando a consistência na formatação das sentenças textuais;
- **Remoção de pontuação:** Pontuações e caracteres especiais nos relatos são eliminados, simplificando o conteúdo textual e facilitando sua análise;
- **Remoção de acentos e sinais diacríticos:** Acentos e sinais diacríticos presentes nos relatos são removidos, normalizando o texto e reduzindo a possibilidade de inconsistências na representação dos caracteres;
- **Conversão para um *Huggingface dataset*:** Conversão dos dados para um *dataset* compatível com o **modelo BERT**, com atributos renomeados para "text" e "label", em formato *Huggingface dataset*, para garantir a interoperabilidade e a integração adequada dos dados com os modelos e ferramentas utilizados neste trabalho.

Conforme observado na Tabela 3, a execução desta fase implicou em uma redução no número de registros de boletins. Neste contexto, é importante ratificar que tais operações e tratamentos nos dados são essenciais para garantir a qualidade e a consistência dos dados, preparando-os adequadamente para a etapa subsequente de modelagem.

Tabela 3: Classes de crimes mais comuns em Marabá entre 2019 e 2023 após a fase de Limpeza de Dados.

Classes do crime	Registros - Brutos	Registros - Após Limpeza
FURTO	14.974	14.837
ROUBO	9.152	9.104
DANO NO TRÂNSITO	6.289	6.254
AMEAÇA	5.703	5.573
ESTELIONATO	5.605	5.485
LESÃO CORPORAL	3.870	3.740
ACIDENTE DE TRÂNSITO SEM VÍTIMA	2.670	2.617
LESÃO NO TRÂNSITO	1.437	1.368
INVASÃO DE DISPOSITIVO INFORMÁTICO	1.162	1.150
DANO	974	956
TOTAL	51.836	51.084

3.5 Implementação de Modelos BERT

Esta fase compreende a aplicação da ferramenta de modelagem selecionada ao conjunto de dados pré-processados, visando à criação de modelos preditivos para classificação supervisionada de sentenças textuais. No caso deste trabalho, como já está sendo utilizado um modelo pré-treinado BERT, a modelagem compreende principalmente o processo de *fine-tuning* desse *Transformer*.

O processo de *fine-tuning* dos modelos *Hugging Face* consiste na adaptação de modelos de linguagem pré-treinados para tarefas específicas por meio de treinamento adicional em conjuntos de dados anotados. Essa prática é fundamental para otimizar o desempenho do modelo em tarefas específicas, como classificação de texto, sumarização ou tradução, via ajuste dos pesos das camadas do modelo para a nova tarefa-alvo. Isto é, pode-se fazer modificações no modelo, como congelar camadas pré-treinadas, modificar o número de camadas ocultas, a configuração do modelo ou incorporar camadas adicionais conforme necessário para uma tarefa específica.

Além disso, o processo de *fine-tuning* permite que modelos pré-treinados, como o BERT, sejam utilizados de forma eficaz em uma ampla variedade de tarefas de NLP, adaptando-se às características e nuances dos dados específicos da aplicação (Wolf et al., 2020). Assim, os modelos *Hugging Face* podem ser personalizados e ajustados para atender às necessidades específicas do contexto de aplicação, resultando em soluções mais eficientes e precisas

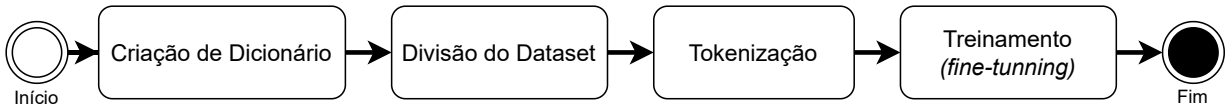

Figura 3: Representação da execução de etapas para implementação de ajuste fino em modelos BERT.

para uma variedade de problemas de NLP.

A Fig. 3 apresenta uma representação da execução dos passos para alcançar o ajuste fino desejado. No primeiro passo da metodologia de treinamento do modelo BERT, destaca-se a criação de um dicionário de correspondência entre os rótulos das classes e seus respectivos identificadores, bem como um mecanismo de retorno dos identificadores para os rótulos originais, a fim de possibilitar o uso eficiente do modelo durante o treinamento e a inferência.

O segundo passo compreende a divisão do conjunto de dados em três partes distintas: 80% para treinamento, 10% para validação e 10% para teste. Essa abordagem permite uma avaliação adequada do desempenho do modelo, bem como a otimização de seus hiperparâmetros, caso seja necessário. É importante destacar que esta divisão do conjunto de dados é realizada de forma aleatória entre os subconjuntos de treinamento, validação e teste. Isso significa que, por exemplo, na segunda vez que esse passo for executado, um registro de um crime que estava no *dataset* de treinamento poderá estar localizado no *dataset* de validação. Este processo é importante para os testes não-paramétricos previstos na fase de Avaliação de Resultados. Além disso, durante este processo de embaralhamento (*shuffle*), os dados são estratificados pelo atributo "label", isto é, as classes dos crimes. Essa estratégia garante a mesma quantidade de registros para cada classe entre os três conjuntos de dados (treinamento, validação e teste).

Em seguida, o terceiro passo contempla a tokenização de todo o conjunto de dados utilizando o modelo BERT pré-treinado Bertimbau ou "neuralmind/bert-base-portuguese-cased" (Souza et al., 2020). Essa etapa é necessária para preparar os dados de entrada para o modelo citado, convertendo-os em sequências de tokens compreensíveis para o modelo.

Finalmente, o quarto passo emprega o uso da classe *AutoModelForSequenceClassification*, disponibilizada pela biblioteca *Hugging Face*, para criar o modelo de classificação de sequência, empregando também o dicionário previamente criado. Durante esta etapa, foram configurados os parâmetros necessários para o treinamento do modelo, incluindo os conjuntos de dados de treinamento, de validação e de teste.

3.6 Avaliação de Resultados

Após a realização do *fine-tuning* do modelo, a fase de Avaliação de Resultados busca estimar o desempenho de classificação dos modelos avaliados. Desta forma, este trabalho considera como métricas de avaliação precisão, *recall*, *F1-score*, acurácia, macroagregado e média ponderada, as quais são comumente utilizadas para avaliar o desempenho de modelos de classificação em tarefas de NLP. Cada uma dessas métricas oferece perspectivas distintas sobre a capacidade do modelo em classificar corretamente as instâncias em diferentes classes, conforme:

- **Precisão** é definida como a proporção de exemplos positivos corretamente classificados em relação ao total de exemplos classificados como positivos. Essa métrica é útil para avaliar a confiabilidade das previsões positivas do modelo, sendo expressa por:

$$\text{Precisão} = \frac{TP}{TP + FP}, \tag{1}$$

  onde TP denota os verdadeiros positivos - *True Positive* (TP) e FP denota os falsos positivos - *False Positive* (FP);
- ***Recall*** representa a proporção de exemplos positivos corretamente identificados em relação ao total de exemplos positivos no conjunto de dados. Esta métrica é especialmente relevante em cenários onde a identificação de todos os exemplos positivos é crucial, sendo representada por:

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{2}$$

  onde FN denota os falsos negativos - *False Negative* (FN);
- **F1-score** é uma medida de desempenho que combina precisão e *recall* em uma única métrica, calculando a média harmônica entre essas duas medidas. Essa métrica é útil para balancear a importância de ambas as métricas e proporcionar uma visão global do desempenho do modelo, conforme:

$$F_1 = 2 \cdot \frac{\text{Precisão} \cdot \text{Recall}}{\text{Precisão} + \text{Recall}}; \tag{3}$$

- **Acurácia** é a proporção de exemplos corretamente classificados em relação ao total de exemplos no conjunto de dados. Embora seja uma métrica comumente utilizada, a acurácia pode ser enganosa em conjuntos de dados desbalanceados, por isso a necessidade de avaliação de outras métricas. Essa métrica é dada por:

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{4}$$

  onde TN denota o total de verdadeiros negativos - *True Negative* (TN);
- **Suporte** refere-se ao número de ocorrências de cada classe no conjunto de dados. Essa métrica fornece informações sobre a distribuição das classes e pode ajudar a identificar classes desbalanceadas que podem impactar a interpretação das métricas de desempenho;

- **Macroagregado e Média Ponderada** calculam a média de cada uma das métricas para cada classe. Enquanto a macroagregada atribui o mesmo peso para todas as classes, a média ponderada considera o peso das classes com base no suporte de cada classe. Assim, é possível obter uma visão geral do desempenho do modelo em todas as classes e identificar padrões de desempenho diferenciados entre as classes.

Outra métrica considerada é a matriz de confusão, que fornece uma representação tabular das relações entre as previsões do modelo e as classes reais dos dados. Assim, a matriz de confusão permite uma análise detalhada do desempenho do modelo em cada classe individualmente, identificando possíveis padrões de erro e áreas de melhoria.

Adicionalmente, a fase de Avaliação de Resultado adota as métricas de *Area Under the Curve* (AUC) e a curva *Receiver Operating Characteristic* (ROC) para avaliação do desempenho de modelos de classificação. A AUC é uma medida da capacidade do modelo em distinguir entre classes positivas e negativas ao variar o limiar de classificação. Essa métrica varia de 0 a 1, onde um valor de 1 indica um modelo perfeito que faz todas as previsões corretas, enquanto um valor próximo a 0,5 indica um modelo que faz previsões aleatórias. Por outro lado, a curva ROC representa graficamente a capacidade de um modelo em distinguir entre classes positivas e negativas, variando os limiares de decisão. Quanto mais a curva ROC se aproxima da região do lado esquerdo superior, maior é a qualidade do teste em relação à sua habilidade de discriminar entre os grupos. Esta análise permite uma avaliação visual do *trade-off* entre sensibilidade e especificidade do modelo para diferentes limiares de classificação.

Por outro lado, conforme previsto no desenho metodológico deste trabalho e com o interesse de avaliar a significância dos resultados obtidos, esta etapa ainda contempla a aplicação do teste estatístico não-paramétrico de Friedman (Friedman, 1937). Este procedimento foi empregado a partir da aplicação dos dados de teste nos modelos avaliados *BERTimbau* e BERT. Este procedimento foi complementado pela aplicação dos métodos *post-hoc* de Shaffer (Shaffer, 1986) e Bergmann (Bergmann and Hommel, 1988).

Neste contexto, o teste de Friedman é um teste de comparações múltiplas utilizado para detectar diferenças de significância entre dois ou mais algoritmos. Na primeira etapa do teste, são consideradas as K_a $(K_a - 1)/2$ possíveis comparações entre os K_a algoritmos. Os resultados dessas comparações podem ser ordenados pelo seu valor p de forma crescente em uma espécie de ranking calculado por cada teste utilizando uma aproximação normal. Esse procedimento considera os seguintes passos:

i. **Reunir os Resultados:** Os resultados observados para cada par de algoritmo e problema são coletados para cada teste realizado;
ii. **Classificação dos Resultados:** Os valores obtidos são classificados de 1 (representando o melhor resultado) a K (representando o pior resultado), conforme o desempenho em relação a cada problema em particular. Essa classificação é denotada como r_{ij}, onde i indica o problema e j representa o algoritmo;
iii. **Cálculo da Média das Classificações:** Para cada algoritmo p, é calculada a média das classificações que foram obtidas em todos os problemas analisados. Esta média, denotada como R_j, é expressa por: $R_j = \frac{1}{n_p} \sum_i r_{ij}$, onde n_p é o número total de problemas.

No contexto do desempenho dos modelos, para cada variante foram realizadas 33 iterações, garantindo maior robustez na análise dos resultados obtidos. Em cada iteração, foi calculado o valor do *F1-score* para cada classe analisada. Posteriormente, as médias de *F1-score* obtidas ao longo das iterações foram calculadas e utilizadas como base para a aplicação dos testes estatísticos não paramétricos. Esse procedimento permitiu a comparação do desempenho entre os modelos, considerando a distribuição dos resultados sem pressupor normalidade nos dados.

4 Resultados e Discussão

Esta seção apresenta os resultados das execuções realizadas com os modelos BERT e Bertimbau. Em particular, esta seção apresenta a avaliação de resultados usando as métricas de acurácia, *recall* e *F1-score*, assim como para as métricas ROC-AUC. De modo complementar, esta seção também apresenta a avaliação realizada por meio da matriz de confusão, além dos resultados do teste estatístico não-paramétrico.

4.1 Avaliação de Desempenho: Acurácia, Recall e F1-score

A Tabela 4 apresenta uma síntese dos resultados obtidos para cada um dos modelos avaliados, considerando as métricas de avaliação aplicadas na predição das classes previstas no dataset. A partir dos resultados obtidos, observa-se que o modelo *BERTimbau* alcançou uma acurácia geral de 92% (0,92), enquanto o modelo BERT alcançou uma acurácia geral de 90% (0,90), ambos para todas as classes e usando os 5.108 registros presentes no conjunto de dados de teste.

Em particular ao modelo *BERTimbau*, é possível verificar por meio da métrica *F1-score*, que os piores desempenhos do modelo foram para as classes de "acidente de trânsito sem vítima" e "dano". Considerando estas classes, o modelo apresentou um resultado de *F1-score* abaixo de 72%. Em uma análise inicial, o motivo para essa baixa eficácia aparenta ter sido causado pela baixa quantidade de amostras (suporte) das classes supracitadas. Por exemplo, a classe "dano" possui apenas 96 elementos para serem usados no teste. Contudo, em análise posterior, é possível comparar os resultados das classes "acidente de trânsito sem vítima" e "invasão de dispositivo informático" com o número de suporte de 262 e 115, respectivamente, mas com valores de *F1-score* notadamente diferentes para cada, sendo 72% (0.72) e 92% (0.92), na mesma ordem. Esta questão indica a necessidade de investigação em outros parâmetros a serem adotados nos modelos a fim de encontrar o motivo e/ou solução para essa diferença no desempenho em classes que possuem quantidades semelhantes de registros.

Tabela 4: Métricas de precisão, *recall* e F1-score no conjunto de dados de teste.

Métrica Classes de crimes	Precisão BERTimbau	Precisão BERT	Recall BERTimbau	Recall BERT	F1-score BERTimbau	F1-score BERT	Suporte
Acidente de Trânsito sem Vítima	0,61	0,61	0,87	0,78	0,72	0,69	262
Ameaça	0,94	0,89	0,92	0,88	0,93	0,89	557
Dano	0,66	0,55	0,56	0,50	0,61	0,52	96
Dano no Trânsito	0,94	0,92	0,76	0,79	0,84	0,85	625
Estelionato	0,94	0,94	0,94	0,93	0,94	0,94	549
Furto	0,96	0,96	0,97	0,95	0,97	0,95	1483
Invasão de Dispositivo Informático	0,96	0,96	0,89	0,90	0,92	0,93	115
Lesão Corporal	0,89	0,81	0,92	0,87	0,91	0,84	374
Lesão no Trânsito	0,80	0,80	0,89	0,83	0,84	0,81	137
Roubo	0,97	0,95	0,97	0,98	0,97	0,96	910
Acurácia					0,92	0,90	5108
Média Macro	0,87	0,84	0,87	0,84	0,86	0,84	5108
Média Ponderada	0,92	0,90	0,92	0,90	0,92	0,90	5108

Em contrapartida, a classificação das demais classes apresentou um desempenho promissor, uma vez que elas demonstraram um índice de *F1-score* acima de 84%. Para essas classes, o suporte é consideravelmente maior, situação que potencialmente contribui com uma eficácia aprimorada.

Além disso, a Tabela 4 demonstra que o desempenho do modelo BERT segue a tendência descrita anteriormente para o modelo *BERTimbau*, mas com algumas ressalvas: as classes "acidente de trânsito sem vítima" e "dano" também apresentaram os piores desempenhos no modelo base, contudo houve uma melhora em ambas as classes usando o modelo treinado na língua portuguesa.

4.2 Avaliação de Desempenho: Métrica ROC-AUC

Seguindo na análise dos resultados, a Fig. 4 apresenta a avaliação da curva ROC para o modelo BERTimbau. Os resultados indicam que o modelo possui uma capacidade satisfatória de discriminação entre as classes. Por exemplo, 8 das 10 classes exibiram uma taxa de acerto viável, apresentando, para isso, um valor acima de 90% (0,90). Por outro lado, as classes de "dano" com AUC = 0,77 e "dano de trânsito" com AUC = 0,87 apresentaram o pior desempenho em relação às demais classes, porém, ainda assim, com resultados satisfatórios para a classificação das classes de crimes.

Desta forma, o desempenho para categorias como "dano" (AUC de 0,77) foi inferior, indicando dificuldades do modelo em diferenciar adequadamente essa classe de outras. Esse resultado pode ser consequência de uma ambiguidade inerente aos dados de entrada ou de um menor volume de dados de treinamento para essa classe. Classes como "dano no trânsito" e "lesão no trânsito" apresentaram valores intermediários de AUC (0,875 e 0,942), sugerindo que, embora o modelo apresente precisão viável, ainda há espaço para melhorias.

Novamente, conforme demonstrado pelas Fig. 4 e Fig. 5, os modelos *BERTimbau* e BERT estão com desempenho próximo, apesar de o modelo treinado na língua portuguesa apresentar rendimento superior em 9 das 10 classes de crime usadas na comparação.

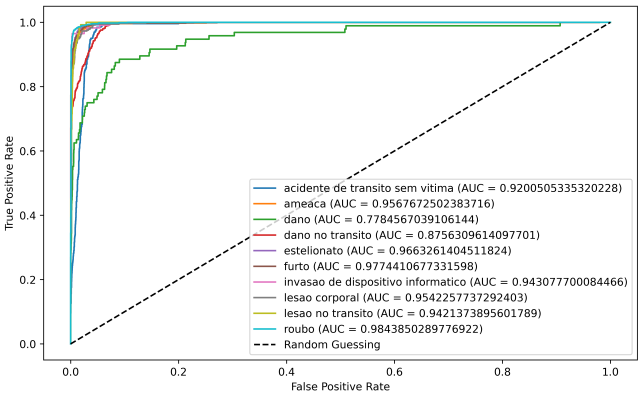

Figura 4: BERTimbau, área sob a curva ROC para as classes de crimes.

4.3 Avaliação de Desempenho: Matriz de Confusão

A Fig. 6 apresenta a matriz de confusão referente ao modelo *BERTimbau*. A matriz demonstra no eixo y as "Classes verdadeiras" e no eixo x exibe como essas "Classes Previstas" classificadas pelo modelo. Pelo mapa de calor nessa matriz, é possível verificar que a classe que o modelo mais fez previsões corretas foi a de "furto", sendo um total de 1439 acertos. Para essa mesma classe, ele fez 24 classificações incorretas como se fossem da classe de "roubo". A análise dessa métrica permitiu uma avaliação abrangente do desempenho do modelo proposto em relação à classificação das diferentes classes, mostrando-se essencial para compreender a capacidade discriminativa e a precisão do modelo.

Por outro lado, a classe "dano" apresentou confusões significativas, com 10 ocorrências sendo classificadas como "furto" e outras 8 como "lesão corporal". Essa dificuldade pode estar associada à generalidade dos termos usados para descrever esses crimes, tornando-os menos distintivos para o modelo. Outro ponto de atenção é a classe "invasão de dispositivo informático", que, embora tenha um desempenho relativamente aceitável, apresenta um número considerável de erros. Casos dessa classe fo-

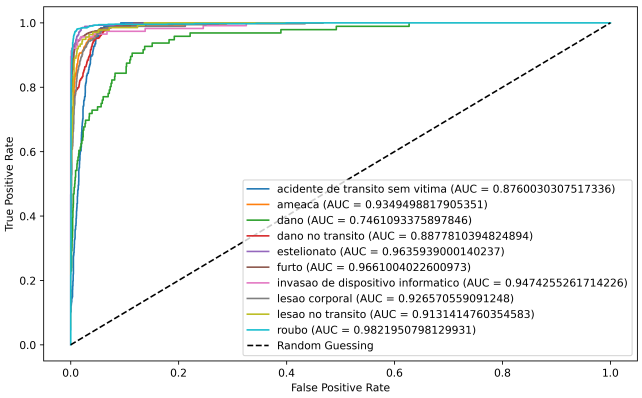

Figura 5: BERT, área sob a curva ROC para as classes de crimes.

ram erroneamente classificados como "estelionato" em 11 instâncias, sugerindo que essas categorias compartilham descritores textuais semelhantes ou que os dados de treinamento não contêm informações suficientes para diferenciá-las adequadamente. Outro caso de sobreposição nas características de tipos de crimes, que é acentuado no gráfico, ocorre entre as classes de "acidente de trânsito sem vítima" e "dano no trânsito". Para esse caso, a classe "dano no trânsito" é classificada erroneamente 130 vezes como se fosse da classe "acidente de trânsito sem vítima".

De modo complementar, a Fig. 7 apresenta a matriz de confusão referente ao modelo BERT, de forma a sintetizar a distribuição de classificações corretas e incorretas entre as classes analisadas. Observa-se que a classe "Furto" apresenta o maior número de predições corretas (1407), evidenciando alta precisão para esta categoria. O mesmo ocorre com a classe "Roubo" que também demonstra bom desempenho com 888 classificações corretas. Por outro lado, erros relevantes foram identificados em categorias como "Dano no trânsito", que foi frequentemente confundida com "Acidente de trânsito sem vítima" (110 casos) e "Lesão no trânsito" (13 casos). Esses resultados sugerem que há desafios na distinção entre categorias relacionadas ao contexto de trânsito, possivelmente devido à similaridade semântica entre as descrições dos eventos.

Por sua vez, a classe "Invasão de dispositivo informático" apresentou um desempenho inferior, com 103 classificações corretas, mas também apresentou classificações incorretas em categorias adjacentes. Essa distribuição reforça a necessidade de ajustes no modelo para lidar com classes menos representadas ou semanticamente próximas.

A comparação entre as matrizes de confusão dos dois modelos revela que ambos apresentam alto desempenho nas categorias mais comuns, como "Furto" e "Roubo". No entanto, o modelo *BERTimbau* demonstra uma ligeira vantagem na classificação de categorias menos representadas, como "Invasão de dispositivo informático" e "Lesão corporal". Essa vantagem pode estar relacionada à capacidade aprimorada de processar texto em português com diferenciação entre letras maiúsculas e minúsculas, contribuindo para uma melhor interpretação semântica. Em contrapartida, ambos os modelos enfrentam desafios na distinção de categorias relacionadas ao trânsito, como "Acidente de trânsito sem vítima" e "Dano no trânsito". Esses resultados sugerem a necessidade de incorporar estratégias adicionais de pré-processamento ou técnicas de aumento de dados para melhorar o reconhecimento dessas categorias específicas.

Além disso, a maior quantidade de classificações corretas observada no modelo *BERTimbau* indica uma possível vantagem na adaptação ao português, tornando-o mais apropriado para aplicações locais. Essa diferença pode ser explicada pela arquitetura do modelo, que foi especificamente treinada em textos em português, ao contrário do modelo BERT base, que possui um treinamento mais genérico em língua inglesa.

4.4 Testes Não-Paramétricos

Para a aplicação do teste de Friedman, os resultados originais de desempenho, expressos em termos da métrica de F1, foram avaliados diretamente, considerando os pressupostos paramétricos do teste. A Tabela 5 sintetiza o *ranking* computado no teste Friedman para cada modelo, destacando o *BERTimbau* como superior ao BERT na classificação. Essas análises permitem inferir que existe significância nas múltiplas comparações possíveis entre as variações citadas do modelo BERT, e que a hipótese H_0 pode ser refutada em favor de H_1.

Tabela 5: Ranking médio computado pelo teste de Friedman.

Modelo	Ranking
BERTimbau	0,99
Bert	1,99

Por outro lado, a Tabela 6 apresenta os valores de p não-ajustado (Friedman) e ajustados pelos procedimentos *post-hoc* de Shaffer e Bergmann. Em todos os casos, os valores ajustados foram idênticos ao valor p não ajustado $(9,22e-9)$, indicando que o ajuste para múltiplas comparações não alterou as conclusões. Essa consistência reforça a confiabilidade do resultado, pois demonstra que as diferenças entre os algoritmos são suficientemente grandes para permanecerem significativas mesmo sob ajustes rigorosos. Isso reforça a rejeição da hipótese nula H_0 e a aceitação da hipótese alternativa H_1.

Tabela 6: Valores p não-ajustado (Friedman) e ajustado (Shaffer e Bergmann) para múltiplas comparações.

Hipótese	p não-ajustado	p Shaffer	p Bermann
BERT vs BERTimbau	9,215887204197207E-9	9,215887204197207E-9	9,215887204197207E-9

Em síntese, os testes estatísticos reforçaram as conclusões obtidas nos experimentos, indicando que o modelo *BERTimbau* apresentou desempenho superior em comparação ao BERT na maioria das análises realizadas. Esses resultados reforçam a robustez do *BERTimbau* em tarefas de processamento de linguagem natural, destacando seu potencial de generalização superior ao do BERT.

Figura 6: Matriz de confusão das classes de crimes - BERTimbau.

Figura 7: Matriz de confusão das classes de crimes - BERT.

5 Considerações Finais

O estudo propôs uma metodologia para a classificação BOPs por meio de técnicas de NLP, utilizando os modelos BERT e *BERTimbau* em uma base textual com diferentes categorias de ocorrências. Para isso, foram aplicadas técnicas de pré-processamento, como normalização, remoção de stopwords e tokenização, visando padronizar os dados utilizados no treinamento dos modelos.

Os resultados indicaram que o modelo *BERTimbau* apresentou desempenho superior ao BERT-base, tanto em termos de acurácia quanto nas análises estatísticas não-paramétricas. A avaliação estatística revelou diferenças significativas, favorecendo o *BERTimbau* em todos os cenários analisados. O modelo alcançou uma acurácia de 92% e um *F1-score* superior a 84% em 8 das 10 classes testadas. Entretanto, algumas previsões incorretas sugerem a necessidade de refinamento do modelo, seja por meio do aumento da representatividade de classes menos frequentes no conjunto de treinamento, seja pela adoção de técnicas mais avançadas de pré-processamento. Além disso, a inclusão de informações contextuais, como horário, local e descrição detalhada das ocorrências, poderia enriquecer o modelo e reduzir a confusão entre categorias.

A análise da matriz de confusão evidenciou erros comuns na classificação, especialmente em categorias semanticamente próximas, como "acidente de trânsito sem vítima" e "dano no trânsito". Esses equívocos ressaltam desafios na captura de nuances entre classes relacionadas, sugerindo a necessidade de refinamento das características utilizadas ou de aprimoramento da diversidade e qualidade dos dados de treinamento. Os testes não-paramétricos reforçaram a robustez dos resultados, indicando que o *BERTimbau* supera o BERT-base no contexto analisado. Os achados ressaltam a importância da aplicação de métodos rigorosos de análise estatística na avaliação do desempenho de algoritmos de aprendizado de máquina, especialmente em problemas complexos como a classificação automatizada de boletins de ocorrência.

Por fim, os experimentos demonstraram que os modelos baseados em BERT, especialmente o *BERTimbau*, são eficazes na identificação de padrões e na classificação de categorias em textos de boletins de ocorrência, mesmo diante de múltiplas classes e variações na distribuição dos dados. Apesar dos bons resultados, ajustes adicionais, como a experimentação de diferentes hiperparâmetros além dos padrões disponibilizados na biblioteca *Hugging Face*, podem contribuir para um desempenho ainda mais robusto. A metodologia proposta se mostra válida e representa uma inovação na aplicação de modelos de linguagem em segurança pública.

Para trabalhos futuros, é relevante explorar configurações mais avançadas dos modelos utilizados, incluindo ajustes finos de hiperparâmetros, aumento no número de épocas de treinamento e avaliação de diferentes estratégias de regularização. Isso poderia fornecer uma visão mais aprofundada sobre o potencial de aprimoramento dos algoritmos no contexto analisado. Além disso, a inclusão de bases de dados maiores e mais diversificadas permitiria avaliar a generalização dos modelos em cenários mais amplos e variados, ampliando o alcance dos resultados obtidos. Deve-se enfatizar, ainda, que a anonimização dos dados utilizados não foi realizada neste trabalho, constituindo uma limitação relevante sob a perspectiva ética e de privacidade. A implementação deste processo em projetos futuros é recomendada, especialmente considerando a natureza sensível das informações presentes nos boletins de ocorrência e a legislação brasileira sobre proteção de dados.

Referências

Barros, T., Pires, C. E. and Filho, D. N. (2021). Sumarização automática de notícias crime no contexto da polícia federal, *Anais Estendidos do XXXVI Simpósio Brasileiro de Bancos de Dados*, SBC, Porto Alegre, RS, Brasil, pp. 127–133. http://dx.doi.org/10.5753/sbbd_estendido.2021.18174.

Barros, T. S. (2022). *Um modelo bert para sumarização extrativa de textos em documentos da polícia federal*, Master's thesis, Universidade Federal de Campina Grande, Campina Grande, Paraíba, Brasil. Available at http://dspace.sti.ufcg.edu.br:8080/jspui/handle/riufcg/27174.

Bergmann, B. and Hommel, G. (1988). Improvements of general multiple test procedures for redundant systems of hypotheses, *Multiple Hypothesenprüfung / Multiple Hypotheses Testing*, Springer Berlin Heidelberg, pp. 100–115. http://dx.doi.org/10.1007/978-3-642-52307-6_8.

Bomfim, T. S. and Lopes, A. d. A. (2022). Mineração de texto aplicada à detecção de cyberbullying. Available at https://bdta.abcd.usp.br/item/003127892.

Brasil (1940). Decreto-lei nº 2.848, de 7 de dezembro de 1940. código penal - cp, Diário Oficial da União, Brasília, DF, 31 dez. 1940. Available at https://www.planalto.gov.br/ccivil_03/decreto-lei/del2848compilado.htm. Acessado em 16/05/2024,.

Brasil (1997). Código de trânsito brasileiro – ctb. lei nº 9.503, de 23 de setembro de 1997. institui o código de trânsito brasileiro, Diário Oficial da União. Available at https://www.planalto.gov.br/ccivil_03/leis/l9503.htm. Acessado em 16/05/2024.

Carneiro, L. d. A. (2022). An essay on violent deaths and criminal incidences in brazil: a descriptive analysis of the national panorama, *Research, Society and Development* **11**(2): e23711225704. http://dx.doi.org/10.33448/rsd-v11i2.25704.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805*. http://dx.doi.org/10.48550/arXiv.1810.04805.

Eufrazio, F. F. (2024). Análise das mortes violentas intencionais de negros/as nordestinos/as pela violência policial, *Revista Brasileira de Segurança Pública* **18**(2): 336–355. http://dx.doi.org/10.31060/rbsp.2024.v18.n2.1906.

FAPESPA (2024). Resultados do Produto Interno Bruto Municipal.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance, *Journal of the American Statistical Association* **32**(200): 675–701. http://dx.doi.org/10.1080/01621459.1937.10503522.

Fórum Brasileiro de Segurança Pública (2024). *18º Anuário Brasileiro de Segurança Pública*, Fórum Brasileiro de Segurança Pública, São Paulo. Available at https://publicacoes.forumseguranca.org.br/handle/123456789/253.

IBGE (2022). IBGE Cidades - Marabá. Available at https://cidades.ibge.gov.br/brasil/pa/maraba/panorama.

Matos, H., Souza, S., Santos, R., Costa, J. and Costa, C. (2022). A supervised classifier for police reports at the state of pará, brazil, *Anais da II Escola Regional de Alto Desempenho Norte 2 e II Escola Regional de Aprendizado de Máquina e Inteligência Artificial Norte 2*, SBC, Porto Alegre, RS, Brasil, pp. 21–24. http://dx.doi.org/10.5753/erad-no2.2022.228238.

Oliveira, M. G. d. (2023). Análise das emoções em redes sociais como ferramenta para estudar manifestações populares, *Trabalho de conclusão de curso (graduação) - Universidade Federal de Santa Maria, Centro de Tecnologia, Curso de Engenharia de Computação, RS* . Available at http://repositorio.ufsm.br/handle/1/30526.

Roy, S. and Goldwasser, D. (2021). Analysis of nuanced stances and sentiment towards entities of US politicians through the lens of moral foundation theory, *in* L.-W. Ku and C.-T. Li (eds), *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, Association for Computational Linguistics, Online, pp. 1–13. http://dx.doi.org/10.18653/v1/2021.socialnlp-1.1.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures, *Journal of the American Statistical Association* **81**(395): 826–831. http://dx.doi.org/10.1080/01621459.1986.10478341.

Silveira, R., Ponte, C., Almeida, V., Pinheiro, V. and Furtado, V. (2023). Legalbert-pt: A pretrained language model for the brazilian portuguese legal domain, *Anais da XII Brazilian Conference on Intelligent Systems*, SBC, Porto Alegre, RS, Brasil, pp. 268–282. Available at https://sol.sbc.org.br/index.php/bracis/article/view/28420.

Souza, F., Nogueira, R. and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese, *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*. http://dx.doi.org/10.1007/978-3-030-61377-8_28.

Souza, S. L. d. (2022). *Mineração de dados em bancos de dados de segurança pública no estado do pará, brasil*, Dissertação de mestrado, Universidade Federal do Pará, Belém. http://dx.doi.org/10.5753/erad-no2.2022.228247.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q. and Rush, A. (2020). Transformers: State-of-the-art natural language processing, *in* Q. Liu and D. Schlangen (eds), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Association for Computational Linguistics, Online, pp. 38–45. http://dx.doi.org/10.18653/v1/2020.emnlp-demos.6.

Yu, B., Tang, F., Ergu, D., Zeng, R., Ma, B. and Liu, F. (2024). Efficient classification of malicious urls: M-bert - a modified bert variant for enhanced semantic understanding, *IEEE Access* **PP**: 1–1. http://dx.doi.org/10.1109/ACCESS.2024.3357095.