Comparative Analysis of Balancing Techniques with Synthetic Data in the Classification of Police Reports

Luan Silva[1], **Gustavo Henrique**[2], **Leandro Araújo**[1], **Adam Santos**[2] **Anderson Soares**[1]

[1]Instituto de Informática
Universidade Federal de Goiás (UFG)
Campus Samambaia – Alameda Palmeiras, s/n – Chácaras Califórnia -
Goiânia – GO, BR – Brazil

[2]Faculdade de Sistemas de Informação
Universidade Federal do Sul e Sudeste do Pará (UNIFESSPA) – Marabá, PA – Brazil

luan.silva@discente.ufg.br, gustavobroechl25@gmail.com,
leandronetlink@gmail.com, adamdreyton@unifesspa.edu.br,
andersonsoares@ufg.br

Abstract. *This paper investigates data balancing techniques for classifying police reports from Altamira/PA (2019–2023) using BERTimbau fine-tuned on 22,117 records across ten criminal categories. We compare Back-Translation, Paraphrasing, Class Weights, and Synthetic Generation via LLM (Sabiá 7B) on a dataset with imbalance ratio of up to 25.4:1. While majority classes achieve F1-scores between 0.93 and 0.96 without balancing, Synthetic Generation via LLM yields the best overall results (Macro F1: 0.896, Weighted F1: 0.911), demonstrating that text generation strategies effectively improve model robustness in imbalanced NLP tasks within sensitive domains.*

Resumo. *Este trabalho investiga técnicas de balanceamento de dados para classificação de boletins de ocorrência de Altamira/PA (2019–2023), utilizando o BERTimbau em ajuste fino sobre 22.117 registros em dez classes criminais. Comparam-se Back-Translation, Paraphrasing, Class Weights e Geração Sintética via LLM (Sabiá 7B) em um conjunto com razão de desbalanceamento de até 25,4:1. Embora as classes majoritárias alcancem F1-score entre 0,93 e 0,96 sem balanceamento, a Geração Sintética via LLM obteve os melhores resultados globais (F1 Macro: 0,896; Weighted: 0,911), demonstrando que estratégias de geração de texto são eficazes para melhorar a robustez do modelo em PLN com dados desbalanceados em domínios sensíveis.*

1. Introdução

A inteligência artificial (IA) tem ganhado destaque e notoriedade em diversas facetas da nossa sociedade, tornando-se indispensável para tarefas do cotidiano. O avanço do aprendizado de máquina (AM) tem promovido transformações significativas em áreas críticas como segurança pública, mercado financeiro, medicina diagnóstica e afins; ao possibilitar a análise de grandes volumes de dados e a identificação de padrões relevantes destes dados, para a tomada de decisão mais assertiva. Modelos baseados em transformadores, como o BERT, destacam-se pela capacidade de capturar representações contextuais e se adaptar a tarefas específicas via fine-tuning [Devlin et al. 2019], sendo amplamente aplicados em PLN sobre grandes volumes de dados textuais. Nesse contexto, no domínio da

segurança pública, destaca-se a aplicação de técnicas de Processamento de Linguagem Natural (PLN) na análise de boletins de ocorrência (BOs), documentos que registram eventos criminais ou administrativos e que frequentemente contêm informações sensíveis, como nomes, endereços e documentos pessoais [Souza et al. 2025].

O desbalanceamento de dados é um problema recorrente em tarefas de PLN, caracterizado pela distribuição desigual entre as classes de um conjunto de dados. Nesse cenário, modelos de aprendizado de máquina tendem a favorecer as classes majoritárias, resultando em alto desempenho global, mas baixa capacidade de generalização para classes minoritárias. Esse comportamento é especialmente crítico em aplicações sensíveis, como análise de documentos legais, detecção de fraudes e classificação de ocorrências, onde eventos raros podem ter maior relevância[Souza et al. 2025].

Na literatura podemos encontrar abordagens que demonstram a aplicação de técnicas de balanceamento de dados com diferentes idiomas (inglês, alemão, azerbaijano e afins), tarefas de classificação para texto (classificação de tópicos, análise de sentimentos, detecção de linguagem ofensiva etc.) e bases de dados diversas [Kesgin and Amasyali 2024, Bencke and Moreira 2024, Zhao et al. 2023, Zhang et al. 2024, Ziyaden et al. 2024, Cauas et al. 2025, de Castro Oliveira et al. 2025].

Este estudo tem como objetivo demonstrar uma análise comparativa de técnicas de balanceamento de dados para o domínio de segurança pública, aplicado à classificação de boletins de ocorrência, com foco nas seguintes técnicas: *Back-Translation*, *Paraphrasing*, *Class Weights* e Geração Sintética via *Large Language Models* (LLM). A base de dados escolhida consiste em dados coletados junto à Secretaria Adjunta de Inteligência e Análise Criminal (SIAC), vinculada à Secretaria de Segurança Pública e Defesa Social (SEGUP) do estado do Pará, Brasil, sendo composta por dez categorias criminais, com maior incidência em boletins de ocorrência de um município pré-definido, não restringido apenas à esse recorte.

De acordo com o Anuário Brasileiro de Segurança Pública, disponibilizado pelo Fórum Brasileiro de Segurança Pública (FBSP), o município de Altamira/PA apresenta uma das maiores taxas do país em crimes relacionados a mortes violentas intencionais e estupro [Brasil 2024, Silva et al. 2024]. Em razão desses elevados índices, Altamira foi selecionada como objeto de estudo para a coleta dos dados criminais.

As demais seções deste trabalho estão organizadas da seguinte forma: a Seção 2 apresenta a fundamentação teórica relativa ao domínio; a Seção 3 descreve os trabalhos relacionados a aplicação de técnicas de balanceamento de dados em PLN; a seção 4 descreve detalhadamente a metodologia adotada para o trabalho; a Seção 5 expõe a análise comparativa das técnicas implementadas, bem como os resultados alcançados; e, por fim, a Seção 6 discute as conclusões referentes aos experimentos e aponta direções para pesquisas futuras.

2. Fundamentação teórica

Esta seção apresenta os fundamentos teóricos da metodologia adotada, abordando técnicas de balanceamento de dados em PLN, o uso de aprendizado de máquina em linguagem natural e as métricas de avaliação.

Com o avanço do aprendizado profundo, modelos baseados em transformadores,

como o BERT, destacaram-se pela capacidade de capturar representações contextuais da linguagem e se adaptar eficientemente a tarefas específicas por meio de fine-tuning [Devlin et al. 2019], impulsionando soluções mais robustas para análise de grandes volumes de dados textuais [Souza et al. 2025]

2.1. Técnicas de balanceamento de dados

O desbalanceamento de classes é um problema sério em PLN, pois a distribuição das categorias-alvo é quase sempre distorcida em tarefas deste tipo[Henning et al. 2023]. Pode-se definir desbalanceamento de classes como um cenário de classificação no qual uma ou mais classes (classes minoritárias) são consideravelmente menos frequentes do que outras (classes majoritárias). Definições mais concretas, por exemplo, relativas à proporção relativa até a qual uma classe é vista como minoritária, dependem da tarefa, da base de de dados[Henning et al. 2023]. Os métodos de balanceamento podem ser categorizados de acordo com sua base: amostragem, aumento de dados, escolha da função de perda, aprendizado em etapas ou projeto do modelo[Wei and Zou 2019, Yang et al. 2020, Buda et al. 2018]. Estudos recentes investigam tanto métodos tradicionais, como *back-translation*, quanto abordagens mais recentes baseadas em modelos de linguagem de grande porte (LLMs).

3. Trabalhos relacionados

O Estado da arte sobre a aplicação de técnicas de balanceamento de dados em PLN apresenta uma ampla variedade de abordagens, deste técnicas de reamostragem, como geração de dados sintéticos, além de testar funções perda diferentes. Nesta seção, destacam-se cinco estudos com sinergia ao tema deste trabalho.

No contexto de cenários com poucos dados, Bencke e Moreira [Bencke and Moreira 2024] investigam o uso de múltiplas técnicas de aumento de dados, como *back-translation*, parafraseamento e geração textual. Os experimentos, realizados com modelos baseados em BERT, demonstram ganhos expressivos de até 43% na métrica F1 em tarefas de classificação de textos, incluindo aplicações em português no domínio de cidades inteligentes. Apesar dos resultados promissores, os autores destacam que algumas técnicas podem comprometer a preservação semântica dos dados gerados, evidenciando um desafio importante no uso de DA.

Com o avanço dos LLMs, Zhao et al. [Zhao et al. 2023] exploram o uso do ChatGPT para geração de dados sintéticos voltados à classificação textual. Os resultados mostram melhorias significativas, especialmente com o modelo BERT, cujo F1 macro aumentou de 49,87% para 65,73%. O estudo também analisa fatores como tamanho e diversidade das amostras geradas, demonstrando que maior variabilidade resulta em melhor desempenho. Entretanto, a avaliação é limitada a um único conjunto de dados (Reuters), e questões relacionadas ao controle de qualidade das amostras geradas não são discutidas em profundidade.

Outros estudos reforçam a eficácia de diferentes abordagens de balanceamento: [Kesgin and Amasyali 2024] propõem o *Modified Cyclical Curriculum Learning* (MCCL), combinando dados reais e sintéticos de forma estratégica em tarefas como análise de sentimentos e detecção de linguagem ofensiva; [Ziyaden et al. 2024] aplicam tradução automática para expandir bases em línguas de baixo recurso; e

[Zhang et al. 2024] propõem uma técnica inspirada no SMOTE combinada com engenharia de prompts, obtendo melhorias consistentes em precisão, *recall* e F1.

De forma geral, observa-se que, embora existam avanços significativos no uso de dados sintéticos para classificação de textos, ainda há lacunas importantes relacionadas à preservação semântica, à avaliação em cenários reais e à aplicação em domínios sensíveis. Em particular, poucos trabalhos investigam o uso dessas técnicas em bases altamente desbalanceadas e compostas por dados sensíveis, como boletins de ocorrência. Nesse contexto, este trabalho busca contribuir ao avaliar comparativamente diferentes estratégias de balanceamento com dados sintéticos, considerando tanto o desempenho dos modelos quanto a qualidade das amostras geradas.

4. Metodologia

O presente estudo adota uma abordagem metodológica baseada no CRISP-DM, com adaptações voltadas ao desenvolvimento de um classificador de textos aplicado a relatos de boletins de ocorrência, bem como à análise comparativa de técnicas de balanceamento de classes no domínio de PLN. A metodologia CRISP-DM é um modelo de processo comumente usado em projetos que envolvem mineração de dados, cujas fases englobam: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implementação. Apesar de a metodologia possuir etapas bem definidas, o modelo permite flexibilidade e pode ser adaptado conforme as necessidades do projeto.

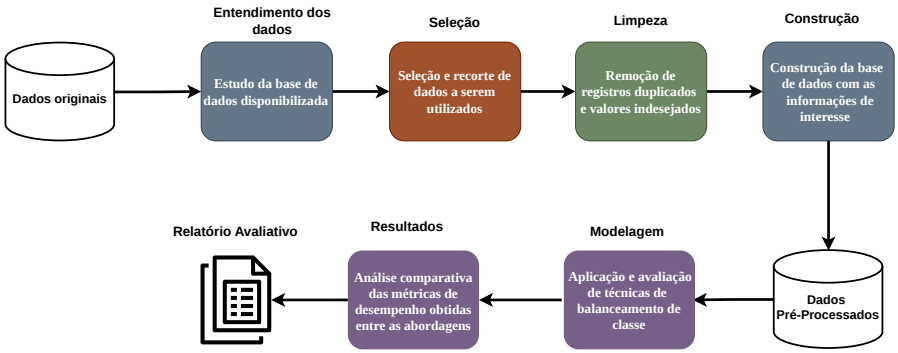


Figure 1. Fases adaptadas do Modelo CRISP-DM

4.1. Entendimento dos Dados

A base de dados utilizada neste trabalho engloba uma série de boletins de ocorrência (BO) disponibilizados pela Secretaria Adjunta de Inteligência e Análise Criminal (SIAC) do estado do Pará, coletados no período de 2019 a 2023. Os boletins de ocorrência contêm textos com informações sobre vítimas, autores, localização, *modus operandi*, entre outros elementos do evento registrado. É importante mencionar que os campos disponibilizados para preenchimento no sistema não possuem validação padronizada ou correção automática, o que pode levar à inserção de registros imprecisos, especialmente em casos de emergência.

A padronização das classes criminais é realizada por analistas da SIAC com conhecimento da legislação brasileira, tornando o campo ”consolidado” ideal para ferramentas de apoio à classificação automática em relatórios de segurança pública.

4.2. Seleção dos Dados

A base de dados original possui um total de 2.400.792 amostras e 72 campos, (e.g,. Data, município, idade), por se tratar de um grande conjunto de dados, foi necessário um recorte geográfico para o município de Altamira que apresenta um desbalanceamento entre ocorrências. Conforme ilustrado na Figura 2, a classe majoritária (Furto) possui 25,4 vezes mais amostras que a classe minoritária (Recuperação de veículos). Este desbalanceamento motivou a investigação de técnicas de balanceamento descritas na seção 4.4.

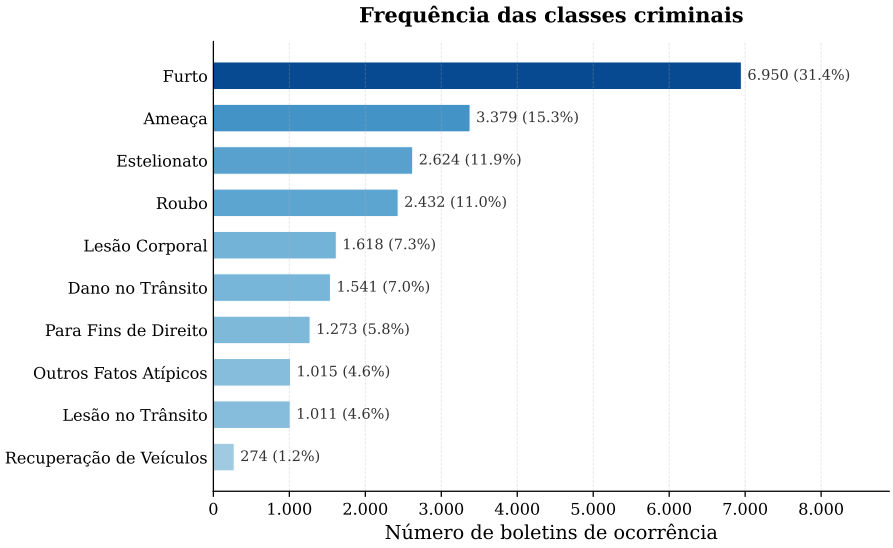


Figure 2. Frequência das classes criminais no período de 2019 a 2023 na cidade de Altamira/PA

4.3. Limpeza e Construção

Com base no recorte realizado para Altamira/PA, o conjunto de dados diminuiu de 2.400.792 amostras para 37909, uma redução significativa e necessária para viabilizar a metodologia escolhida. No entanto, esses dados possuem conteúdo indesejado, que pode prejudicar os resultados dos algoritmos. Com base nisso, é necessário que os dados passem por um processamento que viabilize o treinamento do modelo e o uso das técnicas supramencionadas. São descritas, a seguir, as etapas realizadas:

- **Caracteres especiais e *stopwords***: A fim de simplificar o conteúdo textual da ocorrência, *tags* HTML provenientes da página web foram removidas. Acentos, pontuação e sinais também entraram no filtro de remoção, além de *stopwords*, i.e., palavras como artigos e preposições que não possuem relevância semântica para a classificação.
- **Amostras duplicadas**: Utilizou-se a função drop-duplicates da biblioteca Pandas na coluna de relato, após o tratamento de caracteres especiais, a fim de remover amostras duplicadas que não são interessantes para o treinamento.
- **Informação sensível**: A fim de evitar que dados sensíveis sejam expostos, como vítimas, dados pessoais, localidades e endereços, foi criada uma função *regex* que substitui esses dados por etiquetas (e.g., RG, CPF e LOCAL) para preservar a estrutura do texto sem expor informações sensíveis.

Além disso, todos os caracteres textuais foram convertidos para minúsculos. Ao final desta etapa, restaram 22117 ocorrências prontas para serem utilizadas pelos algoritmos.

4.4. Modelagem

Com o intuito de identificar a melhor abordagem para a problemática das classes desbalanceadas, foram exploradas quatro técnicas de balanceamento de dados para PLN distintos, nos quais são: *Back-Translation*, *Paraphrasing*, *Class Weights* e Geração Sintética via LLM.

As técnicas de balanceamento possuem suas particularidades, no qual é necessário algumas definições, nas quais são:

- **Back-Translation** A técnica de back-translation é aplicada às seis classes minoritárias do conjunto de dados — Recuperação de Veículos, Lesão no Trânsito, Outros Fatos Atípicos, Para Fins de Direito, Dano no Trânsito e Lesão Corporal — com o objetivo de elevá-las ao limiar de 2000 amostras. O procedimento consiste na tradução automática dos textos do português para o inglês, seguida de retradução para o português, explorando as variações lexicais e sintáticas introduzidas no processo como mecanismo de diversificação textual.Para essa etapa, é utilizado o modelo M2M100 (*facebook/m2m100_418M*),que permite tradução multilíngue direta em um único modelo, simplificando o pipeline de geração. Ao final, são gerados 5268 textos sintéticos, resultando em um conjunto balanceado com 27385 amostras.
- ***Paraphrasing*** A técnica de *paraphrasing* segue a mesma estratégia de balanceamento aplicada às classes minoritárias, com o objetivo de atingir o limiar de 2000 amostras por classe. O procedimento consiste na geração de reformulações textuais que preservam o significado semântico original, ao mesmo tempo em que introduzem variações sintáticas e lexicais. Para isso, utiliza-se o modelo Sabiá 7B [Pires et al. 2023], um modelo de linguagem (LLM) desenvolvido para o português brasileiro, configurado com amostragem estocástica e limite de 256 tokens, esse modelo foi escolhido por ser especificamente pré-treinado em português brasileiro, o que o torna mais adequado ao domínio linguístico dos boletins de ocorrência em comparação a modelos multilíngues de propósito geral, como LLaMA ou GPT-4. Cada texto original é utilizado como entrada para geração de uma nova versão parafraseada. Ao final do processo, são produzidos 5268 novos textos, totalizando 27385 amostras no conjunto balanceado.
- **Class Weights** A abordagem de *class weights* consiste na reponderação das classes durante o treinamento do modelo, sem a geração de novos dados. Os pesos são calculados de forma inversamente proporcional à frequência de cada classe no conjunto de treino, com o objetivo de penalizar mais fortemente os erros em classes minoritárias. A implementação é realizada por meio da customização da função de perda, utilizando entropia cruzada ponderada no treinamento com a biblioteca Transformers. Nessa abordagem, a base de dados permanece inalterada, totalizando 22117 amostras.
- **Geração Sintética via LLM** A abordagem de geração sintética via modelos de linguagem de grande porte (LLM) é aplicada a todas as classes com menos de 2000 amostras. Diferentemente das abordagens anteriores, essa técnica utiliza

few-shot prompting para gerar novos exemplos textuais a partir de referências reais da mesma classe. O modelo Sabiá 7B é empregado nesse processo, utilizando de 2 a 3 exemplos reais como contexto no prompt, com variação da *seed* para garantir diversidade. São adotados critérios mínimos de qualidade, como textos com pelo menos 50 caracteres e 15 palavras. Ao final, seis classes minoritárias são elevadas ao limiar de 2000 amostras, enquanto as demais permanecem inalteradas, resultando em um conjunto final com 27385 registros.

O limiar de 2000 amostras foi definido como ponto intermediário entre as classes majoritárias e minoritárias (Figura 2), visando reduzir o desbalanceamento sem promover super-representação artificial das classes minoritárias.

As técnicas de *Back-Translation*, *Paraphrasing* e Geração Sintética via LLM adicionaram 5268 amostras cada, totalizando 27385 registros (+23,82%), enquanto *Class Weights* manteve os 22117 originais. A geração sintética foi realizada exclusivamente sobre as amostras de treino, após divisão estratificada, evitando vazamento de informações do conjunto de teste.

Cada abordagem é avaliada por meio de métricas de desempenho obtidas a partir do *fine-tuning* do modelo BERTimbau (*neuralmind/bert-base-portuguese-cased*), um modelo pré-treinado em um extenso corpus de textos em português, o que o torna particularmente adequado ao domínio linguístico dos boletins de ocorrência analisados.

O classificador BERTimbau foi avaliado por meio de validação cruzada estratificada com $k = 10$ *folds*, treinando em 90% dos dados e avaliando nos 10% restantes a cada iteração. A estratificação garante a preservação da proporção das classes em cada partição. Todas as métricas são reportadas como média $\pm$ desvio padrão dos dez *folds*. A validação cruzada foi aplicada sobre o *dataset* já balanceado de cada estratégia, dado o custo computacional proibitivo de regenerar os dados sintéticos para cada *fold*.

Os hiperparâmetros seguiram as recomendações originais do BERT [Devlin et al. 2019]: batch size 16 (treino) e 32 (avaliação), 3 épocas, garantindo comparação justa entre abordagens.

Table 1. ROUGE médio por classe para cada técnica de aumento de dados.

Classe	***Back-Translation***	**Geração Sintética via LLM**	***Paraphrasing***
Lesão no Trânsito	0,43	0,15	0,34
Outros Fatos Atípicos	0,32	0,08	0,34
Para Fins de Direito	0,36	0,09	0,31
Recuperação de Veículos	0,42	0,09	0,38
Média Geral	**0,386**	**0,104**	**0,422**

A fim de caracterizar a diversidade léxica dos textos gerados por cada técnica de augmentação, foi calculado o ROUGE médio por classe entre as amostras geradas e os textos originais de cada categoria. Embora todas as seis classes minoritárias tenham sido aumentadas, a Tabela 1 concentra-se nas quatro classes de menor volume original, nas quais o efeito da augmentação sobre a diversidade léxica é mais crítico; as classes Lesão Corporal e Dano no Trânsito foram omitidas por apresentarem volume intermediário e desempenho já elevado e estável em todas as estratégias (Tabela 2). O ROUGE médio

entre amostras geradas e originais foi utilizado como indicador de diversidade léxica (1), onde valores próximos de 0 indicam maior distanciamento dos textos originais.

4.4.1. Métricas de Avaliação

Para avaliar o classificador, utilizaram-se *Precision*, *Recall* e *F1-Score* agregados nas formas Macro (peso igual entre classes) e *Weighted* (ponderado pelo suporte), sendo o F1 Macro particularmente relevante em cenários desbalanceados, pois evita que métricas globais mascarem o baixo desempenho nas classes minoritárias [Sokolova and Lapalme 2009, Schütze et al. 2008].

5. Resultados

Esta seção apresenta os resultados obtidos a partir da aplicação das técnicas de balanceamento de dados. Os resultados são apresentados por meio de métricas de desempenho do classificador, contemplando acurácia, F1-score (macro e ponderado), precisão e recall para cada configuração experimental. As tabelas e figuras subsequentes detalham o desempenho por classe de ocorrência, evidenciando o impacto das técnicas de balanceamento especialmente nas classes minoritárias da base de dados.

5.1. Discussão

A Tabela 2 apresenta os resultados obtidos no fine-tuning do modelo BERTimbau em cada abordagem de balanceamento, e a Figura 3 sintetiza o desempenho global em termos de métricas Macro e Weighted. De forma geral, observa-se um padrão consistente entre classes majoritárias e minoritárias, com ganho progressivo das técnicas de augmentação sobre o baseline — destaque para a Geração Sintética via LLM, que alcança F1-Score Macro de 0,896 e Weighted de 0,911, os maiores valores entre todas as estratégias avaliadas.

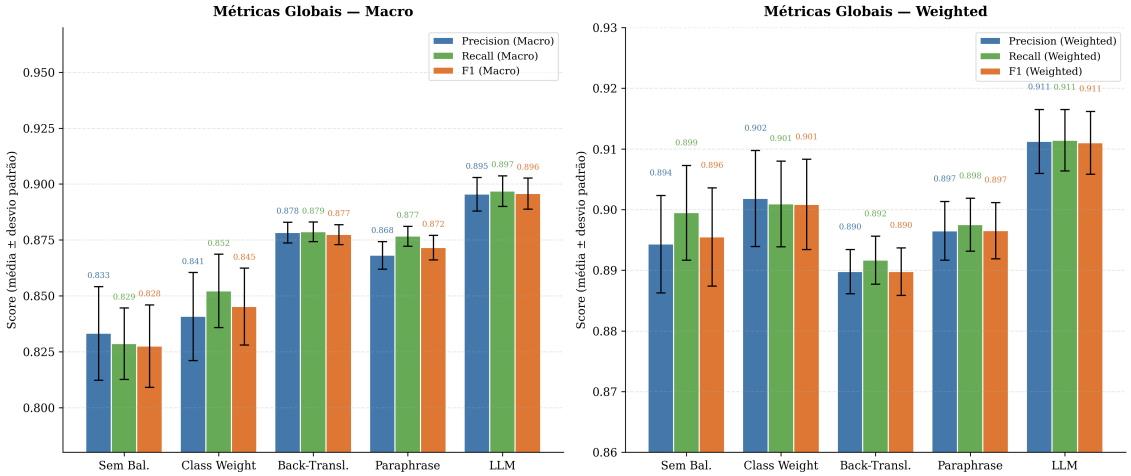


Figure 3. Métricas globais (Macro e Weighted) por estratégia de balanceamento, com média e desvio padrão.

Classes majoritárias como Furto e Roubo, bem como classes com volume intermediário como Dano e Lesão no Trânsito, mantiveram F1-Score alto e estável em todos

os cenários avaliados, variando de 0,93 a 0,96. Esse comportamento era esperado, uma vez que o volume de amostras originais já é suficiente para que o modelo generalize adequadamente, independentemente da estratégia de balanceamento adotada. A estabilidade desses valores também sugere que as técnicas de augmentação não introduziram ruído capaz de prejudicar o desempenho nessas classes.

O cenário mais relevante para a avaliação das técnicas é o das classes minoritárias, onde os efeitos do desbalanceamento são evidentes. Classes como "Para Fins de Direito" e "Outros Fatos Atípicos" apresentaram, no baseline sem balanceamento, F1-Scores de 0,34 e 0,62, respectivamente — valores que indicam dificuldade significativa do modelo em aprender representações confiáveis para essas categorias. Nesse contexto, todas as abordagens produziram ganhos, porém em magnitudes distintas.

Table 2. Resultados das métricas de avaliação de cada técnica de balanceamento utilizada, para cada classe.

Classe	*Precision*					*F1-Score*					*Recall*				
	SB	CW	BT	Par	Sint	SB	CW	BT	Par	Sint	SB	CW	BT	Par	Sint
Furto	0,96	**0,97**	0,95	0,96	0,96	0,96	0,96	0,95	**0,96**	0,96	**0,96**	0,95	0,96	0,96	0,95
Ameaça	0,92	0,91	0,91	**0,92**	0,91	**0,92**	0,91	0,91	0,91	0,91	0,92	0,90	**0,92**	0,91	0,91
Estelionato	0,91	0,92	0,92	0,93	**0,93**	0,92	0,92	0,93	0,92	**0,93**	0,93	0,92	**0,93**	0,92	0,93
Roubo	0,95	0,93	0,95	0,95	**0,95**	0,96	0,95	0,95	**0,96**	0,95	0,97	**0,97**	0,95	0,96	0,96
Lesão Corporal	0,85	0,85	0,81	**0,87**	0,86	0,88	0,87	0,83	**0,88**	0,87	0,90	0,90	0,86	**0,91**	0,89
Dano no Trânsito	0,95	**0,97**	0,94	0,95	0,95	0,95	**0,96**	0,94	0,95	0,95	0,95	**0,96**	0,95	0,95	0,95
Para Fins de Direito	0,46	0,47	**0,72**	0,65	0,69	0,34	0,47	0,66	0,61	**0,69**	0,27	0,48	0,60	0,58	**0,68**
Outros Fatos Atípicos	0,60	0,67	0,74	0,65	**0,80**	0,62	0,66	0,73	0,66	**0,77**	0,65	0,64	0,71	0,66	**0,75**
Lesão no Trânsito	0,93	0,93	0,93	0,92	**0,95**	0,95	0,95	0,93	0,94	**0,96**	0,96	0,97	0,93	0,96	**0,97**
Recuperação de Veículos	0,80	0,78	0,92	0,89	**0,96**	0,79	0,81	0,94	0,92	**0,97**	0,79	0,84	0,97	0,97	**0,98**

Legenda: SB – Sem Balanceamento; CW – *Class Weight*; BT – *Back-Translation*; Par – *Paraphrasing*; Sint – Sintético (LLM).

A abordagem *Class Weights*, por não gerar novos dados mas sim redistribuir a influência das classes durante o treinamento, produziu ganhos modestos e irregulares. Em "Para Fins de Direito", o ganho foi de apenas 0,05 no F1-Score (de 0,53 para 0,58), e embora tenha apresentado resultado mais expressivo em "Outros Fatos Atípicos" (de 0,39 para 0,62), sua eficácia é limitada quando o desequilíbrio é severo, pois o modelo continua exposto a uma distribuição de exemplos enviesada, sem diversificação do espaço de representação. Esse comportamento é visível na Figura 4, em que a matriz do Class Weight ainda apresenta confusão acentuada entre "Para Fins de Direito" e "Outros Fatos Atípicos" (20% e 14%, respectivamente).

A técnica *Back-Translation* apresentou ganhos moderados em "Para Fins de Direito" (de 0,53 para 0,65) e expressivos em "Outros Fatos Atípicos" (de 0,39 para 0,74). Uma hipótese para o desempenho irregular entre as classes está relacionada à natureza não estruturada dos textos de boletins de ocorrência, que apresentam escrita informal, variações ortográficas, abreviações e construções sintáticas irregulares. O processo de tradução e retradução tende a normalizar essas características, produzindo amostras mais padronizadas do que os textos originais, o que pode ter impactado de forma distinta cada classe a depender de sua variabilidade linguística intrínseca.

O *Paraphrasing*, apesar de ser uma técnica semanticamente controlada, apresentou ganhos inferiores aos da *Back-Translation* nas classes mais críticas, elevando o F1-

Score de ”Para Fins de Direito” de 0,53 para 0,62 e de ”Outros Fatos Atípicos” de 0,39 para 0,71. Esse comportamento sugere que, quando o volume de amostras originais é reduzido, o modelo de paráfrase tende a reproduzir as mesmas estruturas linguísticas presentes nos poucos exemplos disponíveis, limitando a diversidade das instâncias geradas e, consequentemente, o ganho no desempenho do classificador.

A Geração Sintética via LLM foi a abordagem que apresentou os ganhos mais expressivos e consistentes nas classes minoritárias, elevando o F1-Score de ”Para Fins de Direito” de 0,53 para 0,71 e de ”Outros Fatos Atípicos” de 0,39 para 0,77 — os maiores valores obtidos entre todas as técnicas avaliadas para essas classes. Esses resultados sugerem que o uso de *few-shot prompting* permitiu ao modelo generativo produzir amostras com maior diversidade semântica, enriquecendo o espaço de representação justamente nas categorias com menor volume de dados originais. Essa interpretação é reforçada pelos scores ROUGE da Tabela 1: com valores de apenas 0,09 em ”Para Fins de Direito” e 0,08 em ”Outros Fatos Atípicos”, os textos gerados pelo LLM se distanciam consideravelmente dos originais em termos léxicos, o que reflete a diversidade semântica que beneficiou a classificação nessas categorias.

A análise conjunta das matrizes de confusão *baseline* e Geração sintética com LLM, apresentadas na Figura 4, reforça as observações anteriores e revela o padrão dominante de erro do classificador: a confusão mútua entre ”Para Fins de Direito” e ”Outros Fatos Atípicos”, que persiste em todas as estratégias mas decresce progressivamente conforme se avança das técnicas baseadas em redistribuição (Class Weight) para as baseadas em augmentação semântica (LLM). Essa convergência indica que o limite de desempenho nessas classes está associado não apenas à escassez de dados, mas também à proximidade semântica entre as próprias categorias — sugerindo que ganhos adicionais dependeriam tanto de mais dados quanto de uma redefinição dos critérios de fronteira entre essas classes na anotação original.

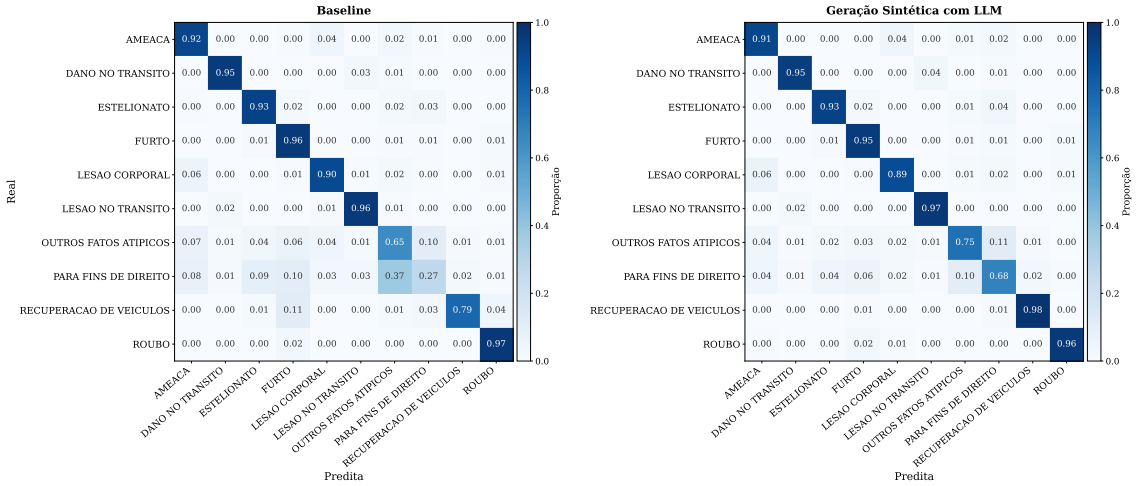


Figure 4. Matriz de Confusão

6. Conclusão e trabalhos futuros

Os resultados demonstram que o modelo BERTimbau apresentou elevado desempenho nas classes majoritárias mesmo na ausência de estratégias de balanceamento, eviden-

ciando sua capacidade de generalização em cenários com ampla disponibilidade de dados. Entretanto, verificou-se que o desbalanceamento compromete significativamente a classificação das classes minoritárias, reforçando a necessidade de técnicas específicas para mitigação desse problema.

Entre as abordagens avaliadas, a geração sintética de dados por meio de LLMs destacou-se como a estratégia mais eficaz, promovendo ganhos substanciais de desempenho, especialmente nas classes minoritárias, além de melhorias consistentes nas métricas analisadas. As técnicas de *Class Weights*, *Paraphrasing* e *Back-Translation* também contribuíram positivamente, ainda que de forma mais moderada, indicando que tanto ajustes na função de custo quanto métodos de reformulação textual auxiliam no aprendizado do modelo em cenários desbalanceados.

Como perspectivas futuras, recomenda-se a investigação de abordagens híbridas de balanceamento, a utilização de modelos mais robustos, teste de significância estatistica, discussão do custo computacional (tempoo de geração de dados sintéticos, consumo de RAM e GPU, e assim por diante) e estratégias de aprendizado contínuo, bem como a realização de análises qualitativas dos dados sintéticos gerados. Adicionalmente, sugere-se a ampliação do estudo para diferentes bases e contextos regionais, visando avaliar a capacidade de generalização dos resultados obtidos em diferentes domínios.

References

Bencke, L. and Moreira, V. P. (2024). Data augmentation strategies to improve text classification: a use case in smart cities. *Language Resources and Evaluation*, 58(2):659–694.

Brasil (2024). Anuário brasileiro de segurança pública 2024. `https://apidspace.universilab.com.br/server/api/core/bitstreams/80177eeb-4a88-40f6-98f5-c476dea0f3db/content`. Acesso em: 01 abr. 2026.

Buda, M., Maki, A., and Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259.

Cauas, P., Pastor, R., de Mattos Neto, P. S., and Duarte, F. C. d. L. (2025). Enhancing crime forecasting with hybrid approaches: Integrating machine learning and statistical models. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 297–308. SBC.

de Castro Oliveira, J. V., Rufino, N., Levcovitz, S., Utino, M. Y. R., Lobato, F., Turine, M., Rezende, S., Junior, A. F. L. J., and Marcacini, R. M. (2025). Agents4gov: Privacy-preserving web agents using open llms for public sector tasks. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 580–591. SBC.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Henning, S., Beluch, W., Fraser, A., and Friedrich, A. (2023). A survey of methods for addressing class imbalance in deep-learning based natural language processing. In

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 523–540.

Kesgin, H. T. and Amasyali, M. F. (2024). Advancing nlp models with strategic text augmentation: A comprehensive study of augmentation methods and curriculum strategies. *Natural Language Processing Journal*, 7:100071.

Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023). *Sabiá: Portuguese Large Language Models*, page 226–240. Springer Nature Switzerland.

Schütze, H., Manning, C. D., and Raghavan, P. (2008). *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Silva, G., Kuribayashi, H., Santos, R., and Santos, A. (2024). Automating the classification of crime reports in altamira, pará using bert-based architectures. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 823–833. SBC.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437.

Souza, V., Santos, A., Soares, A., et al. (2025). Sensitive data protection in police reports: De-identification techniques and applications in machine learning. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 105–116. SBC.

Wei, J. and Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 6382–6388.

Yang, W., Li, J., Fukumoto, F., and Ye, Y. (2020). Hscnn: A hybrid-siamese convolutional neural network for extremely imbalanced multi-label text classification. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 6716–6722.

Zhang, D., Mi, R., Zhou, P., Jin, D., Zhang, M., and Song, T. (2024). Large model-based data augmentation for imbalanced text classification. In *2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, pages 1006–1010. IEEE.

Zhao, H., Chen, H., and Yoon, H.-J. (2023). Enhancing text classification models with generative ai-aided data augmentation. In *2023 IEEE International Conference On Artificial Intelligence Testing (AITest)*, pages 138–145. IEEE.

Ziyaden, A., Yelenov, A., Hajiyev, F., Rustamov, S., and Pak, A. (2024). Text data augmentation and pre-trained language model for enhancing text classification of low-resource languages. *PeerJ Computer Science*, 10:e1974.