Reconhecimento de entidades sensíveis em boletins de ocorrência com modelos baseados na arquitetura transformer

Victor Souza [1], Luan Silva[2], Adam Santos[3], Reginaldo Filho[4], Anderson Soares[2]

[1]Núcleo de Desenvolvimento Amazônico em Engenharia (NDAE)
Universidade Federal do Pará (UFPA)
Rodovia BR-422, Km 13 - Vila Permanente, Tucuruí, PA, Brasil

[2]Universidade Federal de Goiás (UFG)
Campus Samambaia - Alameda Palmeiras, s/n - Chácaras Califórnia, Goiânia, GO, Brasil

[3]Universidade Federal do Sul e Sudeste do Pará (UNIFESSPA)
Folha 17, Quadra 04, Lote Especial, Nova Marabá, PA, Brasil

[4]Universidade do Pará (UFPA)
R. Augusto Corrêa, 01 - Guamá, Belém, PA, Brasil

victor.ferreira.souza@tucurui.ufpa.br, luan.silva@discente.ufg.br

adamdreyton@unifesspa.edu.br, regicsf@ufpa.br, andersonsoares@ufg.br

Abstract. *This work explores the automatic identification of sensitive data in Brazilian police reports using transformer-based deep learning models. Real records from Marabá (PA) were annotated with entities such as* `PESSOA`*,* `ENDEREÇO`*, and* `CPF`*. Portuguese models (BERTimbau, TUCANO, Albertina, and robertalexpt-base) were fine-tuned and evaluated. The* `robertalexpt-base` *model achieved the best F1-score (0.83), confirming the potential of contextualized Portuguese models for privacy-preserving text de-identification.*

Resumo. *Este trabalho aborda a identificação automática de dados sensíveis em boletins de ocorrência brasileiros com modelos de aprendizado profundo baseados em transformers. Registros reais de Marabá (PA) foram anotados com entidades como* `PESSOA`*,* `ENDEREÇO` *e* `CPF`*. Modelos em português foram ajustados e avaliados, destacando-se o* `robertalexpt-base` *com o F1-score (0,83), o que confirma o potencial de modelos contextualizados para desidentificação de textos com preservação de privacidade.*

1. Introdução

A proteção de dados sensíveis tornou-se uma preocupação central na era digital, impulsionada pelo crescimento do volume de informações processadas por sistemas inteligentes. Dados pessoais — como nomes, endereços e identificadores — são amplamente utilizados em aplicações de aprendizado de máquina (AM), o que traz desafios éticos e legais relacionados à privacidade e à segurança da informação [Union 2016, Brasil 2018].

O AM tem promovido avanços em diversas áreas, inclusive na segurança pública, permitindo a análise automatizada de grandes volumes de texto. Nesse contexto, técnicas de processamento de linguagem natural (PLN) vêm sendo aplicadas à análise de boletins

de ocorrência (BOs), que contêm informações textuais sensíveis sobre vítimas, suspeitos e locais. Embora essa análise possa auxiliar na formulação de políticas e na gestão de recursos, ela também impõe riscos de exposição de dados pessoais.

Métodos de reconhecimento de entidades nomeadas (*Named Entity Recognition*, NER) têm sido amplamente empregados para identificar e classificar informações sensíveis [Wang et al. 2023], mas sua precisão ainda é limitada em textos ruidosos e contextos específicos. Diante disso, este trabalho analisa técnicas de identificação de dados sensíveis em BOs, discutindo suas limitações e evidenciando a possibilidade de abordagens automatizadas que busca equilibrar privacidade, segurança e utilidade da informação.

2. Metodologia

A metodologia proposta é composta por três etapas principais: (i) aquisição de dados e anotação dos dados sensíveis, (ii) treinamento e avaliação de modelos de identificação, e (iii) análise dos resultados obtidos.

Na primeira etapa, foram utilizados BOs registrados na cidade de Marabá (PA) entre 2019 e 2023. O estudo concentrou-se nos dez tipos de registros mais frequentes, como furto, roubo e estelionato, que representam a maior parte das ocorrências. A anotação das entidades sensíveis foi conduzida de forma semiautomática, empregando o modelo *GPT-4o-mini*, da OpenAI [OpenAI 2024], configurado para reconhecer categorias predefinidas, como `PESSOA`, `ENDEREÇO`, `CPF`, `VEÍCULO`, entre outras. Em seguida, foi realizada uma revisão manual com o objetivo de corrigir inconsistências e padronizar as anotações. Os dados foram estruturados no formato BIO (*Begin–Inside–Outside*), permitindo que cada token fosse rotulado conforme sua posição dentro das entidades.

Na segunda etapa, foi realizado o *fine-tuning* dos modelos BERTimbau Base, BERTimbau Large, TUCANO 160, TUCANO 630 e robertalexpt-base, todos voltados para a tarefa de NER. Esses modelos, baseados na arquitetura de *transformers*, foram treinados com o mesmo conjunto anotado, utilizando a divisão de 80% dos dados para treinamento, 10% para validação e 10% para teste.

Por fim, os modelos foram avaliados com base em métricas estabelecidas — precisão, revocação, F1-score — a fim de comparar o desempenho entre as abordagens e analisar sua eficácia na identificação automática de dados sensíveis em textos de boletins de ocorrência.

3. Resultados

3.1. Distribuição de tokens anotados

A Tabela 1 apresenta a distribuição final de *tokens* por entidade, incluindo a classe “O”, referente aos tokens fora de entidades nomeadas.

Observa-se um desbalanceamento significativo entre as classes, com destaque para a entidade “PESSOA”, seguida por “ENDEREÇO” e “VEÍCULO”. A classe “O”, com mais de 8 milhões de tokens, representa um padrão comum em tarefas de NER, nas quais apenas parte do texto corresponde a entidades de interesse [Sang and De Meulder 2003]. Essa assimetria reforça a importância do uso de métricas macro e de técnicas de balanceamento para mitigar vieses de aprendizado.

Tabela 1. Distribuição de *tokens* por entidade.

Entidade	Tokens	Entidade	Tokens
PESSOA	365.568	TELEFONE	36.079
BANCO	16.256	CNPJ	2.676
ENDEREÇO	105.735	CNH	2.252
VEÍCULO	87.594	RG	5.356
CPF	15.051	EMAIL	14.426
EMPRESA	38.518	O	8.242.814

3.2. Desempenho geral dos modelos e por entidade

A Tabela 2 apresenta o desempenho global dos modelos, desconsiderando a classe "O", permitindo uma avaliação mais fiel das entidades sensíveis. Todos os modelos baseados em *transformers* apresentaram bom desempenho, com *F1-score* acima de 0,77. O `robertalexpt-base` obteve o melhor resultado médio (0,83), seguido por `BERTimbau Base` e `Albertina 100m`, destacando-se pela robustez na identificação de entidades em português.

Os modelos maiores, como o BERTimbau Large e o TUCANO 630m, não apresentaram ganhos relevantes em relação às versões menores, indicando que o aumento de parâmetros, no intervalo avaliado, não gerou melhorias significativas. Estudos futuros podem investigar modelos ainda maiores para verificar se essa tendência se mantém.

Tabela 2. Desempenho global dos modelos sem considerar a classe O.

Modelo	Acurácia	Macro precisão	Macro revocação	Macro *F1-score*
BERTimbau Base	82%	0,91	0,74	0,81
BERTimbau Large	81%	0,91	0,73	0,81
TUCANO 160m	76%	0,90	0,68	0,77
TUCANO 630m	77%	0,91	0,69	0,78
robertalexpt-base	82%	0,91	0,77	**0,83**
Albertina 100m	81%	0,91	0,73	0,80

Tabela 3. Resultados por entidade (*F1-score*) em diferentes modelos de NER, sem considerar a classe O.

Modelo	BAN	CNH	CNPJ	CPF	EM	EMP	END	PES	RG	TEL	VEI
BERTimbau Base	0,76	0,88	0,98	0,94	0,92	0,80	0,82	0,96	0,90	0,93	0,81
BERTimbau Large	0,76	0,88	0,97	0,94	0,92	0,80	0,82	0,96	0,90	0,93	0,80
TUCANO 160m	0,70	0,80	0,97	0,92	0,91	0,74	0,76	0,95	0,87	0,91	0,74
TUCANO 630m	0,70	0,81	0,97	0,93	0,92	0,75	0,78	0,95	0,86	0,91	0,74
robertalexpt-base	0,81	0,89	0,98	0,95	0,97	0,81	0,83	0,96	0,91	0,94	0,83
Albertina 100m	0,72	0,86	**0,99**	0,94	0,95	0,81	**0,84**	0,96	0,83	0,94	0,81

A Tabela 3 mostra que os modelos obtiveram alto desempenho em entidades com padrões regulares, como `CPF`, `CNPJ` e `TELEFONE`. Em contrapartida, classes mais

ambíguas e contextuais — como ENDEREÇO, EMPRESA e BANCO — apresentaram maior variação. Nessas, o robertalexpt-base destacou-se pelo melhor equilíbrio entre precisão e revocação.

4. Conclusões

Os resultados obtidos demonstram a alta eficácia dos modelos baseados em *transformers* para a identificação automática de dados sensíveis em boletins de ocorrência. Todos os modelos analisados apresentaram desempenho consistente, com valores de *F1-score* acima de 0,77, mesmo diante do forte desbalanceamento entre as classes.

Entre as arquiteturas avaliadas, o robertalexpt-base apresentou o melhor desempenho médio (*F1-score* de 0,83) e maior estabilidade entre as entidades, demonstrando robustez no reconhecimento de padrões contextuais em português. Os modelos BERTimbau Base e Albertina 100m também obtiveram resultados sólidos, destacando-se como opções leves e eficientes.

O desempenho superior do robertalexpt-base está associado aos corpora LegalPT e CrawlPT, ricos em textos jurídicos e dados web em português brasileiro, que ampliaram sua familiaridade com as estruturas e expressões típicas dos boletins de ocorrência.

Observou-se ainda que o aumento do número de parâmetros nas variantes de maior porte, como BERTimbau Large e TUCANO 630m, não resultou em ganhos significativos de desempenho, sugerindo que modelos de tamanho intermediário são suficientes para capturar a complexidade linguística e semântica dos textos analisados, mantendo boa relação entre custo computacional e acurácia.

Entidades com estrutura linguística regular, como CPF, CNPJ e TELEFONE, apresentaram desempenho elevado em todos os modelos, enquanto categorias mais ambíguas e dependentes de contexto — como ENDEREÇO, EMPRESA e BANCO — continuaram sendo os principais desafios para a tarefa de NER.

De forma geral, os resultados reforçam o potencial das arquiteturas baseadas em *transformers* para aplicações em desidentificação de textos em português, demonstrando capacidade de generalização, estabilidade e precisão superiores. Esses achados indicam que abordagens contextualizadas e pré-treinadas no idioma são fundamentais para o avanço de sistemas automáticos de proteção de dados sensíveis no contexto de boletins de ocorrência.

Referências

Brasil (2018). Lei geral de proteção de dados pessoais. Lei n.º 13.709/2018.

OpenAI (2024). Gpt-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence. Acessado em: 25 maio 2025.

Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Union, E. (2016). General data protection regulation. Regulation (EU) 2016/679.

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., and Wang, G. (2023). Gpt-ner: Named entity recognition via large language models.