Sensitive Data Protection in Police Reports: De-identification Techniques and Applications in Machine Learning

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

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

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

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

victor.ferreira.souza@tucurui.ufpa.br, adamdreyton@unifesspa.edu.br

regicsf@ufpa.br, andersonsoares@ufg.br

Abstract. *This paper presents a methodology for identifying and de-identifying sensitive information in police reports using Named Entity Recognition (NER). Two models are evaluated: BERTimbau, a transformer-based model trained in Brazilian Portuguese, and BiLSTM, a traditional recurrent architecture. Experimental results showed that BERTimbau outperformed BiLSTM in macro F1-score and in the precision of de-identification tasks, particularly for minority classes. The study highlights the importance of adopting contextual models and robust evaluation strategies to ensure privacy preservation in unstructured public security data.*

Resumo. *Este trabalho propõe uma metodologia para identificação e desidentificação de dados sensíveis em boletins de ocorrência por meio de técnicas de reconhecimento de entidades nomeadas (NER). São comparados dois modelos: o BERTimbau, baseado em transformers e treinado em português brasileiro, e o BiLSTM, com arquitetura recorrente tradicional. Os resultados indicaram que o BERTimbau obteve desempenho superior em F1-score macro e maior eficácia na desidentificação, especialmente em entidades minoritárias. O estudo reforça a necessidade de modelos contextuais e métricas robustas para garantir a privacidade em dados de segurança pública.*

1. Introdução

O avanço do aprendizado de máquina (AM) tem impulsionado transformações em áreas críticas como saúde, segurança, finanças e serviços públicos, viabilizando a análise de grandes volumes de dados e a descoberta de padrões relevantes para a tomada de decisão [Esteva et al. 2017, Topol 2019, LeCun et al. 2015, Domingos 2012]. 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 contêm informações sensíveis, como nomes, endereços e documentos pessoais [Souza et al. 2022].

A manipulação computacional desses dados, embora promissora para a formulação de políticas públicas e o apoio à investigação, levanta preocupações éticas e legais relacionadas à privacidade dos indivíduos. Em especial, a exposição de identificadores pessoais em sistemas automatizados pode resultar em riscos de reidentificação, contrariando regulamentações como a Lei Geral de Proteção de Dados (LGPD) e o Regulamento Geral de Proteção de Dados da União Europeia (GDPR) [Dwork e Roth 2014, Brasil 2018, Union 2016].

A identificação de dados sensíveis em grandes corpora textuais é fundamental para mitigar tais riscos. Métodos de reconhecimento de entidades nomeadas (NER) são amplamente utilizados para localizar informações como nomes, telefones e endereços [Wang et al. 2023]. No entanto, esses métodos ainda enfrentam limitações importantes em contextos multilíngues, ambíguos ou culturalmente específicos, como os boletins de ocorrência redigidos em português brasileiro [Wang et al. 2020, Yermilov et al. 2023].

Nesse cenário, a desidentificação surge como uma etapa essencial para a preservação da privacidade. Técnicas como pseudonimização e anonimização visam remover ou substituir dados identificáveis, permitindo a reutilização segura dos documentos para fins científicos e operacionais. Além da proteção individual, a disponibilização responsável de dados desidentificados contribui diretamente para o desenvolvimento de modelos de AM em contextos de interesse social, como segurança pública e justiça [Topol 2019]. Para isso, é fundamental equilibrar a utilidade analítica dos dados com a conformidade legal e os princípios éticos.

Este estudo propõe e avalia uma metodologia de identificação e desidentificação de informações sensíveis em boletins de ocorrência, com foco na comparação entre dois modelos de NER: BERTimbau, baseado em *transformers*, e BiLSTM, de arquitetura recorrente. A eficácia dos modelos é analisada em termos de precisão, revocação, *F1-score* e impacto na desidentificação automática. Os resultados contribuem para o uso seguro de dados sensíveis em aplicações de AM voltadas à segurança pública.

2. Fundamentação Teórica

Esta seção apresenta os fundamentos teóricos da metodologia adotada, abordando a identificação e desidentificação de dados sensíveis, o uso de aprendizado de máquina em linguagem natural e as arquiteturas BiLSTM e BERTimbau.

2.1. Identificação e Desidentificação de Dados Sensíveis

A identificação e a desidentificação de dados são etapas fundamentais para garantir a privacidade em sistemas que processam informações textuais sensíveis, como boletins de ocorrência. Tais processos contribuem com legislações como a LGPD [Brasil 2018] e o GDPR [Union 2016], e são especialmente relevantes em contextos nos quais há grande volume de texto não estruturado.

A identificação de dados sensíveis consiste em localizar e classificar elementos como nomes, documentos, endereços e telefones. Para isso, são amplamente utiliza-

das técnicas de NER, baseadas em regras, léxico ou aprendizado de máquina. Modelos modernos como BiLSTM-CRF e BERT têm se destacado pela alta precisão em tarefas de NER, principalmente em contextos linguísticos complexos [Catelli et al. 2021, Muralitharan e Arumugam 2024, Wang et al. 2020].

Após essa etapa, a desidentificação visa proteger a privacidade substituindo ou removendo as entidades sensíveis. As abordagens mais comuns incluem a pseudonimização — em que dados são trocados por identificadores reversíveis — e a anonimização — onde a associação ao indivíduo é permanentemente eliminada [Yermilov et al. 2023, Ohm 2010].

Esses processos são essenciais para viabilizar o uso ético e seguro de dados em pesquisas e políticas públicas, mantendo a utilidade analítica dos documentos sem comprometer a identidade dos envolvidos [Dwork e Roth 2014].

2.2. Aprendizado de Máquina e Processamento de Linguagem Natural

O AM é um campo da inteligência artificial que permite a construção de modelos capazes de extrair padrões e realizar previsões a partir de dados [Cortes e Vapnik 1995]. Algoritmos como *Support Vector Machine* (SVM), árvores de decisão e redes neurais têm sido aplicados em diversas áreas, incluindo segurança pública e análise de documentos textuais.

Modelos de aprendizado profundo, como *Long Short-Term Memory* (LSTM) e *Convolutional Neural Networks* (CNN), ampliaram significativamente a capacidade de representação e generalização desses sistemas, sendo aplicados com sucesso em tarefas de PLN, como análise sintática, tradução e classificação de textos [LeCun et al. 2015].

2.3. Modelos de Linguagem de Grande Escala

Nos últimos anos, os modelos de linguagem de grande escala (LLMs) tornaram-se centrais no PLN. O BERT [Devlin et al. 2019], por exemplo, introduziu o mecanismo de atenção bidirecional, possibilitando a compreensão profunda do contexto textual. Esses modelos demonstraram desempenho excepcional em tarefas como geração de texto, resumo automático e NER [Bommasani et al. 2021].

Apesar dos avanços, os LLMs trazem riscos à privacidade, pois podem memorizar e reproduzir dados sensíveis extraídos durante o pré-treinamento [Carlini et al. 2021]. Estratégias como filtragem com NER têm sido utilizadas para mitigar esse problema, mas sua eficácia em grandes corpora ainda é limitada [Lehman et al. 2021, Yermilov et al. 2023].

2.4. BERTimbau e BiLSTM

O *BERTimbau* é um modelo de linguagem pré-treinado exclusivamente em português brasileiro, baseado na arquitetura BERT e treinado com corpora diversos, como OSCAR, Wikipédia e textos jurídicos [Souza et al. 2020]. Disponível nas versões Base e Large, o BERTimbau mostrou desempenho superior em tarefas sensíveis ao contexto, incluindo NER, sendo compatível com ferramentas modernas como *Hugging Face Transformers*.

Por sua vez, o *BiLSTM* é uma arquitetura de rede neural recorrente que processa sequências de texto em duas direções, capturando dependências de longo prazo em ambos os sentidos [Hochreiter e Schmidhuber 1997, Schuster e Paliwal 1997]. Apesar de

ser superado pelos *Transformers* em diversos benchmarks, o BiLSTM ainda é útil em ambientes com restrições computacionais ou conjuntos de dados pequenos. É comum sua combinação com camadas CRF para garantir consistência na rotulagem sequencial [Lample et al. 2016, Huang et al. 2015].

3. Trabalhos Relacionados

A literatura sobre identificação e desidentificação de dados sensíveis apresenta uma ampla diversidade de abordagens, variando desde técnicas baseadas em regras até métodos avançados de aprendizado profundo. Nesta seção, destacam-se quatro estudos relevantes, enfatizando também suas limitações para melhor contextualizar o presente trabalho.

Dias et al. [Dias et al. 2020] investigaram abordagens híbridas para detecção de dados sensíveis em português europeu, combinando regras, léxico e aprendizado profundo com BiLSTM. O trabalho obteve *f1-score* de 83,01%, porém enfrenta limitações significativas devido à baixa cobertura dos léxicos utilizados para categorias específicas, como profissões e dados médicos. Adicionalmente, a escassez de corpora anotados para português europeu limita a generalização desses modelos para outros contextos.

Catelli et al. [Catelli et al. 2021] apresentaram uma abordagem robusta para desidentificação clínica via BiLSTM+CRF e *embeddings* contextualizados, alcançando *f1-score* de até 96,1%, superando métodos anteriores sem necessidade de engenharia manual. Contudo, destacam-se limitações quanto à necessidade de validação em outros domínios e questões relacionadas à escalabilidade da arquitetura utilizada.

Yayık et al. [Yayık et al. 2021] utilizaram FastText e BERT para classificar tópicos sensíveis segundo a legislação turca, obtendo um *f1-score* de 94,73% com FastText, preferido pela eficiência computacional. Embora eficaz em ambientes industriais, o trabalho possui limitações importantes por não realizar identificação direta de entidades sensíveis, restringindo sua aplicação em tarefas detalhadas de desidentificação.

Muralitharan e Arumugam [Muralitharan e Arumugam 2024] desenvolveram o algoritmo *Privacy BERT-LSTM*, destacando-se pela combinação de *embeddings* contextuais, processamento sequencial e atenção, com *f1-score* de 85,02% no corpus *SMS Spam Collection*. Embora tenha obtido alta performance, a avaliação limitada a um único corpus e a necessidade de ampliar testes para textos mais complexos constituem limitações importantes para sua generalização.

A Tabela 1 resume as principais características e limitações desses estudos, incluindo também o presente trabalho.

Ao analisar esses estudos, percebe-se que, apesar dos avanços obtidos, persistem desafios importantes relacionados à disponibilidade e qualidade dos corpora anotados, adaptação a diferentes domínios e idiomas, bem como limitações na identificação direta e detalhada de entidades sensíveis. Em contraste, o presente trabalho se diferencia ao aplicar técnicas supervisionadas de NER diretamente sobre BOs escritos em português brasileiro, abrangendo categorias como CPF, RG, endereço e nomes de pessoas.

Diferentemente de abordagens limitadas por regras fixas ou dependentes de classificações temáticas prévias, o método proposto neste trabalho utiliza modelos baseados em redes neurais profundas (BiLSTM e BERTimbau), visando maior generalização e aplicabilidade prática. Tal abordagem preenche uma lacuna relevante na literatura, apro-

Tabela 1. Comparação e limitações de estudos.

Referência	Domínio	Idioma	Técnicas	Limitações
Dias et al. (2020)	Documentos gerais	Português europeu	Híbrido (regras, léxico, BiLSTM)	Baixa cobertura léxica e escassez de corpora anotados
Catelli et al. (2021)	Clínico	Inglês	BiLSTM+CRF, embeddings contextualizados	Necessidade de validação em outros domínios, escalabilidade limitada
Yayık et al. (2021)	Jurídico-industrial	Turco	FastText, BERT	Ausência de identificação direta de entidades sensíveis
Muralitharan e Arumugam (2024)	Genérico	Inglês	BERT, LSTM	Avaliação restrita a um corpus, limitada generalização
Presente trabalho	BOs	Português brasileiro	BiLSTM, BERTimbau	Avaliação limitada ao domínio institucional específico

ximando técnicas avançadas de NER da realidade institucional brasileira, onde a proteção efetiva de dados pessoais é crítica.

4. Metodologia de Pesquisa

A metodologia adotada neste trabalho está estruturada em três etapas principais: (i) anotação dos dados; (ii) treinamento de modelos para identificação automática de informações sensíveis com posterior aplicação de técnicas de desidentificação; e (iii) análise dos resultados. A Figura 1 apresenta uma visão geral do fluxo metodológico.

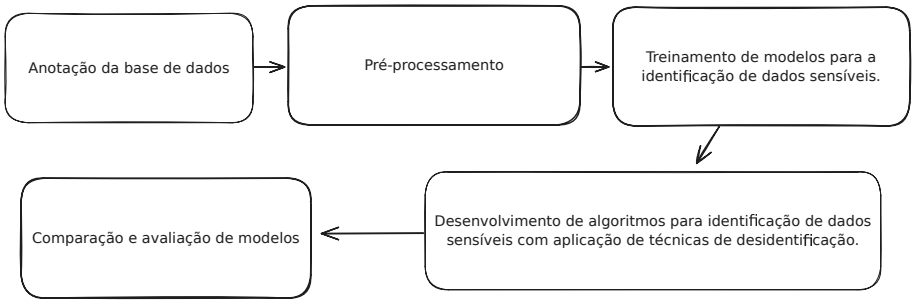

Figura 1. Visão geral da metodologia proposta.

4.1. Anotação dos Dados

A primeira etapa consistiu na seleção e anotação de um corpus textual extraído de boletins de ocorrência registrados na cidade de Marabá, estado do Pará, entre os anos de 2019 e 2023. Foi realizado um recorte das 10 categorias de ocorrência mais frequentes, visando delimitar um conjunto representativo e relevante para a análise.

A anotação inicial foi realizada automaticamente por meio do modelo GPT-4o-mini, da OpenAI, ajustado para identificar e classificar entidades sensíveis com base em categorias previamente definidas [OpenAI 2024]. O processo consistiu na geração de anotações a partir de instruções detalhadas, via *prompt*[1], restringindo a extração às seguintes categorias: `BANCO`, `CNH`, `CPF`, `EMPRESA`, `ENDEREÇO`, `PESSOA`, `RG`, `TELEFONE`, `VEÍCULO`, `CNPJ` e `EMAIL`. O modelo foi acionado por meio da API da OpenAI, com temperatura controlada (0,2) para garantir consistência na saída.

Posteriormente, conduziu-se um processo de revisão manual para correção e refinamento das anotações. Nesta etapa, foram removidas entidades fora do escopo, corrigidas anotações equivocadas e reclassificadas entidades ambíguas. A abordagem híbrida, automatizada e supervisionada, garantiu maior precisão e consistência nas anotações utilizadas nas fases seguintes.

Por fim, o conjunto de dados anotado foi dividido em três subconjuntos: 80% para treinamento, 10% para validação e 10% para teste, assegurando a avaliação imparcial do desempenho dos modelos.

4.2. Treinamento dos Modelos de Identificação

Com os dados anotados, iniciou-se o treinamento dos modelos de NER, com o objetivo de automatizar a identificação de informações sensíveis em textos não estruturados.

Dois modelos com arquiteturas distintas foram utilizados: o BERTimbau Base, baseado em *transformers*, e um modelo clássico com BiLSTM, permitindo uma análise comparativa entre abordagens modernas e tradicionais aplicadas ao PLN em língua portuguesa.

O modelo baseado no BERTimbau foi construído a partir do modelo pré-treinado *neuralmind/bert-base-portuguese-cased*, utilizando a biblioteca *transformers*. A tokenização foi realizada com o algoritmo WordPiece, que fragmenta palavras em subunidades frequentes do vocabulário. Para garantir a consistência entre os *tokens* e os rótulos anotados, foi utilizado o mapeamento de deslocamento de caracteres (`offset_mapping`), e *tokens* especiais como `[CLS]` e `[PAD]` foram ignorados no cálculo da perda, por meio do uso de rótulos `-100`. O comprimento máximo das sequências foi limitado a 512 *tokens*, respeitando a arquitetura do BERT. O treinamento foi conduzido com taxa de aprendizado de 5×10^{-5}, *batch size* de 8, e *weight decay* de 0,01, durante 5 épocas. Foi utilizado o otimizador AdamW com avaliação ao final de cada época, e o melhor modelo foi salvo juntamente com as métricas de desempenho registradas.

Já o modelo BiLSTM foi implementado com a biblioteca *Keras* [Chollet et al. 2015], adotando uma arquitetura com vetores de *embedding* aprendidos do zero e dimensão de 20. A rede contou com uma camada LSTM bidirecional com 50 unidades em cada direção, totalizando 100 unidades ocultas por *token*, o que permitiu capturar dependências contextuais anteriores e posteriores na sequência. Foi aplicado *dropout* recorrente de 0,1 para reduzir o risco de *overfitting*, seguido por camadas densas

[1]Link para o *prompt* utilizado: `https://firebasestorage.googleapis.com/v0/b/sensitive-data-5391d.firebasestorage.app/o/prompt.txt?alt=media&token=b97c1d2c-1aa4-4969-916a-31cab30b9bd9`

com ativações *ReLU* e *softmax* para a classificação dos rótulos. As sentenças foram normalizadas e padronizadas para 512 *tokens*, com os rótulos representados em formato *one-hot*. O modelo foi treinado por 5 épocas, utilizando o otimizador Adam com taxa de aprendizado padrão (1×10^{-3}), *batch size* de 32 e uma divisão de 20% dos dados para validação.

Após o treinamento, os modelos foram avaliados com base nas métricas de precisão, revocação e *F1-Score*, além da geração de matrizes de confusão para análise qualitativa dos acertos e erros [Powers 2020, Manning et al. 2008]. Essa etapa foi fundamental para avaliar a capacidade de generalização de cada arquitetura frente às particularidades linguísticas dos boletins de ocorrência e à variabilidade dos dados sensíveis.

4.3. Desidentificação

Após a identificação das entidades sensíveis, foram aplicadas técnicas de desidentificação com o objetivo de proteger a privacidade dos indivíduos mencionados nos documentos, contribuindo com normas como a LGPD e o GDPR.

A abordagem adotada foi a pseudonimização, na qual as entidades sensíveis foram substituídas por pseudônimos gerados de forma controlada, permitindo uma reidentificação segura em ambientes autorizados [Yermilov et al. 2023].

Essa técnica foi aplicada com base nas saídas dos modelos de NER e avaliada qualitativamente quanto à manutenção do sentido textual e à preservação da utilidade dos documentos desidentificados. Para automatizar a identificação de informações sensíveis, foram empregados os modelos BERTimbau e BiLSTM tratados neste trabalho.

A localização precisa das entidades no texto original exigiu uma etapa de normalização, que incluiu a remoção de acentuação, pontuação e símbolos, além do uso de expressões regulares para alinhar os rótulos às ocorrências reais. Em seguida, realizou-se a pseudonimização, substituindo cada entidade por marcadores genéricos no formato `[TIPO_N]`, conforme seu tipo e ordem de ocorrência.

Como resultado, foram geradas, para cada texto, as versões original, anotada e pseudonimizada, acompanhadas do mapeamento entre as entidades reais e seus respectivos pseudônimos. Esses dados possibilitaram a análise da consistência do processo de desidentificação e da preservação do conteúdo informativo.

5. Resultados e Discussões

Esta seção apresenta os principais resultados da avaliação dos modelos BERTimbau Base e BiLSTM na tarefa de reconhecimento de entidades sensíveis em boletins de ocorrência. Os resultados são discutidos a partir das distribuições de entidades, métricas globais, desempenho por classe e análise qualitativa da desidentificação.

5.1. Distribuição das Entidades

A partir da anotação automatizada e dos pós-processamentos, identificaram-se as entidades sensíveis mais recorrentes, conforme apresentado na Tabela 2. Observa-se a predominância da entidade `PESSOA`, evidenciando a natureza sensível dos textos analisados.

Tabela 2. Quantidade de tokens por entidade identificada

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

5.2. Desempenho Global dos Modelos

A Tabela 3 compara os modelos segundo métricas macro. Apesar do BiLSTM alcançar acurácia superior (99%), o modelo BERTimbau superou em F1-score, precisão e revocação, especialmente por sua capacidade de identificar entidades minoritárias.

Tabela 3. Desempenho global dos modelos BERTimbau e BiLSTM.

Modelo	Acurácia	Macro precisão	Macro revocação	Macro F1-score
BERTimbau Base	96%	0,80	0,81	0,81
BiLSTM	99%	0,73	0,64	0,68

5.3. Desempenho por Entidade

A Tabela 4 evidencia o desempenho de cada modelo por classe. O BERTimbau apresentou vantagem clara na maioria das entidades, com destaque para `PESSOA`, `EMAIL`, `BANCO` e `ENDEREÇO`.

Tabela 4. Métricas por entidade para BERTimbau e BiLSTM.

Entidade	F1-score BERT	F1-score LSTM	Precisão BERT	Precisão LSTM
BANCO	0,67	0,38	0,77	0,67
CNH	0,78	0,65	0,76	0,76
CNPJ	0,87	0,83	0,79	0,88
CPF	0,86	0,74	0,82	0,72
EMAIL	0,87	0,80	0,90	0,79
EMPRESA	0,72	0,63	0,77	0,70
ENDEREÇO	0,70	0,58	0,71	0,65
PESSOA	0,91	0,85	0,90	0,86
RG	0,83	0,77	0,84	0,80
TELEFONE	0,82	0,75	0,77	0,82
VEÍCULO	0,66	0,68	0,64	0,74
O	0,98	0,99	0,98	0,99

5.4. Análise de Matrizes de Confusão

As Figuras 2 e 3 mostram as matrizes de confusão para os modelos BERTimbau e BiLSTM, respectivamente, destacando os erros entre classes principais.

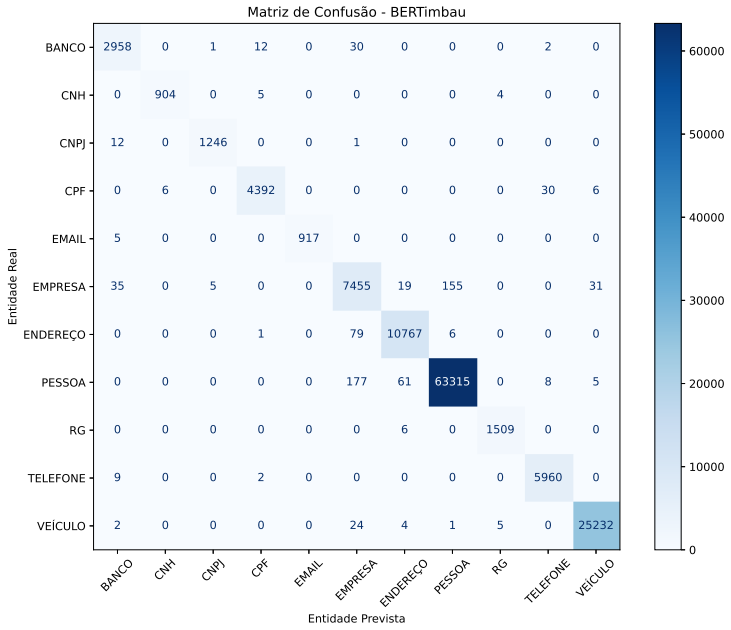

Figura 2. Matriz de confusão — BERTimbau.

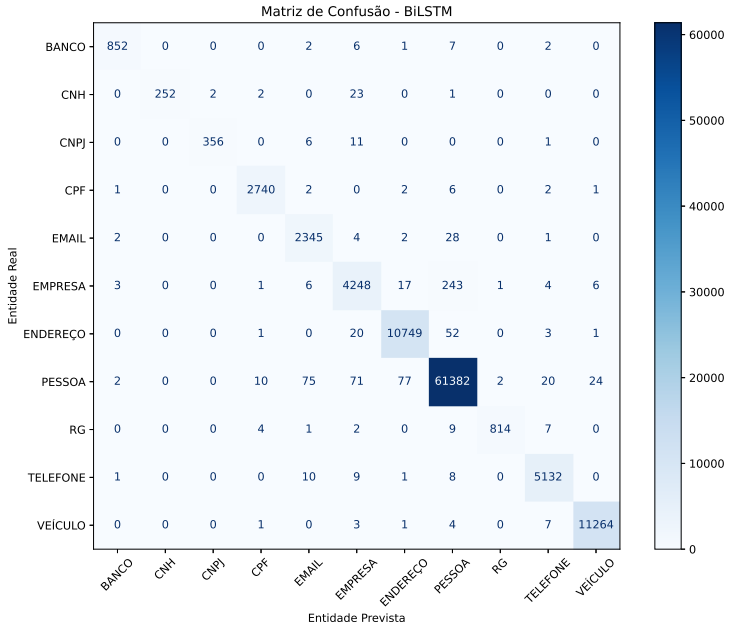

Figura 3. Matriz de confusão — BiLSTM.

A matriz do BERTimbau apresenta menor taxa de confusão entre entidades distintas e maior precisão na separação de classes próximas como `ENDEREÇO` e `EMPRESA`. O

BiLSTM, por outro lado, confunde com mais frequência essas categorias, além de apresentar transições inconsistentes em entidades com baixa frequência.

6. Aplicação de Algoritmos de Desidentificação

Para ilustrar a aplicação prática dos modelos, foram utilizados trechos de boletins de ocorrência fictícios contendo informações sensíveis. As versões dos textos incluem: (i) o original anonimizado manualmente; (ii) a saída do modelo BERTimbau; e (iii) a saída do BiLSTM.

O modelo BERTimbau demonstrou maior capacidade de anonimização, substituindo corretamente entidades como nomes, endereços e veículos por marcadores genéricos (por exemplo, `[PESSOA_1]`, `[VEÍCULO_1]`). Já o modelo BiLSTM apresentou desempenho inferior em alguns casos, deixando parte dos dados sensíveis expostos.

O exemplo abaixo evidencia a relação direta entre a qualidade da etapa de reconhecimento de entidades e a efetividade da desidentificação automatizada.

> **Exemplo:**
> **Texto original:**
> Na data de 15 de março de 2023, JOANA CLARA FERREIRA foi abordada por dois indivíduos armados na AVENIDA MARECHAL RONDON, em frente ao número 1250. Eles levaram seu celular, documentos pessoais e a motocicleta HONDA/BIZ 125 ES, PLACA UEU4H81.
> **Desidentificação com BERTimbau:**
> Na data de 15 de março de 2023, [PESSOA_1] foi abordada por dois indivíduos armados na [ENDEREÇO_1], em frente ao número 1250. Eles levaram seu celular, documentos pessoais e a motocicleta HONDA/BIZ 125 ES, [VEÍCULO_1].
> **Desidentificação com BiLSTM:**
> Na data de 15 de março de 2023, [PESSOA_1] foi abordada por dois indivíduos armados na AVENIDA MARECHAL RONDON, em frente ao número 1250. Eles levaram seu celular, documentos pessoais e a motocicleta HONDA/BIZ 125 ES, PLACA UEU4H81.

7. Conclusão

Os resultados evidenciam a superioridade do BERTimbau na identificação de entidades sensíveis, com melhor desempenho em precisão, revocação e *F1-score*, além de maior robustez a entidades raras. A acurácia de 99% do BiLSTM foi influenciada pelo desbalanceamento do corpus — marcado pela predominância da classe `O` — e não reflete seu desempenho real. Assim, o *F1-score* foi adotado como métrica principal, por equilibrar precisão e revocação e se mostrar mais adequado à aplicação.

O BERTimbau superou o BiLSTM na maioria das categorias críticas, como `PESSOA`, `ENDEREÇO` e `EMPRESA`, com exceção pontual em `VEÍCULO`. Na aplicação prática de desidentificação, o modelo BERTimbau conseguiu anonimizar corretamente os textos, enquanto o BiLSTM deixou informações sensíveis expostas.

Recomenda-se o uso do modelo BERTimbau, ou de arquiteturas semelhantes, em cenários críticos de proteção de dados. Os resultados também destacam a importância de

estratégias de balanceamento de classes, uso de bases mais representativas e avaliações com métricas macro e análises qualitativas.

Como atividades futuras, propõe-se a investigação de modelos mais robustos, como variantes do BERT otimizadas para o português, e o uso de técnicas de *data augmentation* para aumentar a diversidade dos dados anotados. Além disso, pretende-se avaliar se os dados desidentificados mantêm utilidade em tarefas secundárias, como classificação e extração de informações, verificando o equilíbrio entre anonimização eficaz e preservação do valor analítico.

Referências

Bommasani, R., Hudson, D. A., Adeli, E., e et al. (2021). On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

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

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., e Raffel, C. (2021). Extracting training data from large language models.

Catelli, R., Casola, V., De Pietro, G., Fujita, H., e Esposito, M. (2021). Combining contextualized word representation and sub-document level analysis through bi-lstm+crf architecture for clinical de-identification. *Knowledge-Based Systems*, 213:106649.

Chollet, F. et al. (2015). Keras. https://github.com/fchollet/keras.

Cortes, C. e Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Devlin, J., Chang, M.-W., Lee, K., e Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., e Solorio, T., editors, *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, Minneapolis, Minnesota. Association for Computational Linguistics.

Dias, M., Boné, J., Ferreira, J. C., Ribeiro, R., e Maia, R. (2020). Named entity recognition for sensitive data discovery in portuguese. *Applied Sciences*, 10(7).

Domingos, P. (2012). A few useful things to know about machine learning. *Commun. ACM*, 55(10):78–87.

Dwork, C. e Roth, A. (2014). The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., e Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118.

Hochreiter, S. e Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Huang, Z., Xu, W., e Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., e Dyer, C. (2016). Neural architectures for named entity recognition.

LeCun, Y., Bengio, Y., e Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Lehman, E., Jain, S., Pichotta, K., Goldberg, Y., e Wallace, B. C. (2021). Does bert pretrained on clinical notes reveal sensitive data?

Manning, C. D., Raghavan, P., e Schütze, H. (2008). Introduction to information retrieval.

Muralitharan, J. e Arumugam, C. (2024). Privacy bert-lstm: a novel nlp algorithm for sensitive information detection in textual documents. *Neural Computing and Applications*, 36(25):15439–15454.

Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, 57:1701–1777. University of Colorado Law Legal Studies Research Paper No. 9-12.

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.

Powers, D. M. W. (2020). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

Schuster, M. e Paliwal, K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Souza, F., Nogueira, R., e Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Souza, S., Matos, H., Costa, C., Filho, R. S., e Costa, J. (2022). Data mining in public security databases in belém, pará, brazil. In *Anais da II Escola Regional de Alto Desempenho Norte 2 e II Escola Regional de Aprendizado de Máquina e Inteligência Artificial Norte 2*, pages 33–36, Porto Alegre, RS, Brasil. SBC.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56.

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., e Wang, G. (2023). Gpt-ner: Named entity recognition via large language models.

Wang, X., Wang, Z., Han, X., Jiang, W., Han, R., Liu, Z., Li, J., Li, P., Lin, Y., e Zhou, J. (2020). MAVEN: A Massive General Domain Event Detection Dataset. In Webber, B., Cohn, T., He, Y., e Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1652–1671, Online. Association for Computational Linguistics.

Yayık, A., Apik, H., e Tosun, A. (2021). Deep learning based topic classification for sensitivity assignment to personal data. In *2021 6th International Conference on Computer Science and Engineering (UBMK)*, pages 292–297.

Yermilov, O., Raheja, V., e Chernodub, A. (2023). Privacy- and utility-preserving nlp with anonymized data: A case study of pseudonymization.