Avaliação do Reconhecimento de Entidades Nomeadas para Descoberta de Dados Pessoais em Transcrições de Áudio

Carlos André Misiuk Munhos[1], **Luciano Ignaczak**[1]

[1]Universidade do Vale do Rio dos Sinos – 93.022-750 – São Leopoldo – RS – Brasil

carlos.mmunhos@gmail.com, lignaczak@unisinos.br

Abstract. *The growth of personal data have driven legislation such as Brazil's LGPD, but the scarcity of annotated data and the inherently unstructured nature of sources like audio transcripts make both training named entity recognition (NER) models and reliably identifying personal data challenging; to address these issues, this work presents the results of applying a fine-tuned version of the Portuguese pre-trained BERTimbau model with synthetic data to identify four entities (Name, CPF, RG, and Address) and extract the relations "Resides in" and "Holds document" achieving an F1-score of 0.98 in the NER task and 0.29 in the relation extraction task.*

Resumo. *O volume exponencial de dados pessoais impulsionou legislações como a LGPD no Brasil, mas a escassez de dados anotados publicamente e a natureza não estruturada de fontes como transcrições de áudio tornam desafiadores tanto o treinamento de modelos de reconhecimento de entidades nomeadas (NER) quanto a identificação confiável de dados pessoais. Para enfrentar esses desafios, este trabalho apresenta o resultado da aplicação do modelo BERTimbau, refinado com um corpus baseado em dados sintéticos. para identificar quatro entidades (Nome, CPF, RG e Endereço) e extrair as relações "Reside em" e "Possui documento". O modelo refinado alcançou F1-score de 0,98 na tarefa de NER e 0,29 na extração de relações.*

1. Introdução

Dados pessoais são informações que identificam ou podem identificar um indivíduo [Brasil 2018]. Presentes em serviços bancários, saúde e comércio eletrônico, sua proteção é fundamental para garantir privacidade, evitar fraudes e assegurar o controle sobre as próprias informações [Neves 2022]. Diante disso, leis como a GDPR, na Europa, e a LGPD, no Brasil, foram criadas. Vazamentos de dados geram impactos sérios, como danos reputacionais e financeiros [ABNT 2023] e um dos desafios enfrentados pelas organizações é localizar onde seus dados estão armazenados e processados [Gartner 2019], possibilitando assim a implementação dos controles de segurança adequados.

Diversos estudos discutem a aplicação da tarefa de reconhecimento de entidades mencionadas para a identificação de dados pessoais. Neste estudo, utilizaremos a sigla NER, referente ao termo em Inglês *Named Entity Recognition* (NER). Essa tarefa permite analisar e extrair informações de grandes volumes de texto de maneira eficiente, identificando dados como nomes, localizações e outros identificadores pessoais [Aggarwal and Zhai 2012]. Por exemplo, [Moussaoui et al. 2023] empregam NER

para detectar dados pessoais em um corpus com documentos da área jurídica escritos em Árabe, demonstrando a eficácia do método em várias linguagens e contextos. Já [Bannour et al. 2022] menciona que um dos principais desafios para a implementação efetiva dessa tarefa é a escassez de conjuntos de dados anotados disponíveis para treinamento. A qualidade e a abrangência dos dados de treinamento são cruciais para o desempenho desses modelos [Silva et al. 2020], e a falta de acesso a bases de dados ricas e variadas pode limitar significativamente a eficácia do NER na descoberta de dados pessoais.

Pesquisas já publicadas que abordam a descoberta de dados pessoais estão divididas em dois grupos sendo eles o grupo de privacidade e o grupo de desidentificação médica. Nos estudos do grupo de privacidade, [Wongvises et al. 2022], [Herwanto et al. 2021] e [Hu et al. 2022] utilizaram o modelo BERT para implementar a tarefa de NER. Já no contexto médico, [Zhang et al. 2023] implementou o modelo BERT. Por sua vez, o modelo Bi-LSTM+CRF foi utilizado no contexto de privacidade por [Gultiaev and Domashova 2022], enquanto [Catelli et al. 2020] o utilizou em contexto médico. Os estudos selecionados implementaram o modelo em diversos idiomas, como Inglês [Herwanto et al. 2021, Catelli et al. 2021], Tailandês [Wongvises et al. 2022], Chinês [Zhang et al. 2023], Árabe [Moussaoui et al. 2023], Francês [Bannour et al. 2022], Italiano [Catelli et al. 2020] e Russo [Gultiaev and Domashova 2022]. Dessa forma, podemos analisar os resultados obtidos em diferentes contextos linguísticos, avaliando a eficácia das tarefas de NER em múltiplos idiomas.

Este estudo propõe uma abordagem com BERT, utilizando dados sintéticos, para identificar informações pessoais em transcrições de áudio em Português. A pesquisa busca responder: ”Qual é a performance da aplicação das tarefas de NER e extração de relações para a descoberta de dados pessoais em transcrições de áudio em Português?”. Para isso, desenvolveu-se um sistema capaz de identificar as entidades Nome, CPF, RG e Endereço, e também extrair relações entre elas, como ”Reside em”e ”Possui Documento”, com base na LGPD. Utiliza-se o termo NLP (*Natural Language Processing*) para se referir ao Processamento de Linguagem Natural.

Nas seções subsequentes, este artigo aprofundará os trabalhos relacionados e as metodologias empregadas. A Seção 2 evidenciará os estudos relacionados e suas contribuições. Na Seção 3 serão destacadas as metodologias utilizadas para a implementação do experimento. Na Seção 4 são apresentados os resultados obtidos e, por fim, na Seção 5 são avaliados os resultados, limitações e trabalhos futuros.

2. Trabalhos Relacionados

Os trabalhos relacionados apresentam pesquisas e estudos relevantes com o tema proposto, contextualizando-o no âmbito acadêmico. Eles evidenciam as contribuições existentes e lacunas, justificando a relevância deste trabalho. Os artigos escolhidos foram categorizados em dois grupos distintos. O primeiro grupo, denominado ”Privacidade”, engloba os artigos que concentram-se primordialmente na detecção e identificação de dados que estão relacionados à privacidade do titular, e será referido pela sigla ”Priv”. O segundo grupo, intitulado ”Desidentificação Médica”, inclui os artigos que investigam a eficácia dos modelos propostos em realizar a desidentificação médica e será referido pela

sigla ”Med”.

A Tabela 1 demonstra os artigos selecionados, o modelo utilizado e o resultado do F1-score obtido em cada experimento realizado. Analisando primeiramente os artigos do grupo Med, temos diversos modelos utilizados em datasets distintos que estão em diferentes línguas, um problema comum observado foi a escassez de dados para treinamento devido a sua natureza sensível.

Artigo	Ano	F1-score	Modelo	Idioma	N° Entidades	Dataset	Grupo
(BANNOUR et al., 2022)	2022	0.706	Privado	FR	15	MERLOT	Med
(CATELLI et al., 2020)	2020	0.859	Bi-LSTM+CRF	IT	23	SIRM COVID-19	Med
(CATELLI et al., 2021)	2021	0.963	ELECTRA	EN	28	i2b2	Med
(GULTIAEV; DOMASHOVA, 2022)	2022	0.916	BI-LSTM+CRF	RU	2	Nerus	Priv
(HERWANTO; QUIRCHMAYR; TJOA, 2021)	2021	0.728	BERT	EN	2	User Stories Dalpiaz	Priv
(HU et al., 2022)	2022	0.576	KeyBERT	EN	2	Inspec	Priv
(MOUSSAOUI; CHAKIR; BOUMHIDI, 2023)	2023	0.961	Privado	AR	8	Próprio	Priv
(SILVA et al., 2020)	2020	0.860	spaCy	EN	2	Kaggle	Priv
(WONGVISES; KHURAT; NORASET, 2022)	2023	0.573	BERT	TH	7	Próprio	Priv
(ZHANG et al., 2023)	2024	0.915	BERT	CH	8	Próprio	Med

Tabela 1. Resumo Geral dos artigos analisados

Idioma: (EN) Inglês, (TH) Tailandês, (CH) Chinês, (AR) Árabe, (FR) Francês, (IT) Italiano, (RU) Russo.
Modelo: (BERT) Bidirectional Encoder Representations from Transformers, (ELECTRA) “Efficiently Learning an Encoder that Classifies Token Replacements Accurately, (Bi-LSTM) Bidirection Long-Short Term Memory, (CRF) Conditional Random Fields.
Grupo: (Priv) Privacidade, (Med) Desidentificação médica.

Em [Bannour et al. 2022], os autores propuseram uma abordagem baseada em aprendizado profundo para reconhecimento de entidades clínicas em notas narrativas, enfrentando a escassez de anotações por meio de um modelo professor-estudante, que preserva a privacidade e mantém desempenho competitivo (F1-score de 0.706). De forma semelhante, [Catelli et al. 2020] utilizaram estratégias multilíngues para NER clínico, demonstrando que o treinamento cruzado entre Inglês e Italiano pode ser eficaz (F1-score de 0.859). A valorização de recursos linguísticos específicos também apareceu em [Zhang et al. 2023], onde a curadoria de uma base terminológica médica, aliada ao fine-tuning de BERT em textos chineses, permitiu alcançar F1-score de 0.915. Já [Catelli et al. 2021] avançaram na tarefa de desidentificação clínica com o sistema PHI, que incorpora o modelo ELECTRA e agrupamento semântico, obtendo F1-score de 0.963. Esses trabalhos reforçam a importância do uso de modelos contextualizados e da adaptação a recursos linguísticos e contextuais específicos, seja por meio de transferência entre idiomas, curadoria de terminologias ou arquitetura de modelos.

Outros estudos abordam a privacidade em contextos menos explorados ou em línguas com poucos recursos. [Wongvises et al. 2022] investigaram a aplicação de Processamento de Linguagem Natural (PLN) em Tailandês, utilizando o modelo WangchanBERTa para detectar dados pessoais e alcançando F1-score de 0.573. Os autores consideraram o resultado expressivo, considerando o ineditismo da proposta. A proteção de informações pessoais em textos jurídicos árabes foi explorada por [Moussaoui et al. 2023], que propuseram um modelo de NER especializado, alcançando

F1-score de 0.961. Em [Herwanto et al. 2021], o foco do estudo foi na privacidade em histórias de usuários de metodologias ágeis e atingiu um F1-score de 0.728. O estudo de [Gultiaev and Domashova 2022] abordou a desidentificação de dados pessoais em Russo com Bi-LSTM+CRF (F1-score de 0.916), enquanto [Silva et al. 2020] avaliaram as bibliotecas NLTK, Stanford CoreNLP e spaCy com foco em NER aplicado à privacidade. No estudo, a biblioteca spaCy alcançou o melhor resultado (F1-score de 0.860). Por fim, [Hu et al. 2022] ampliaram a discussão ao propor anonimização de dados de fala com técnicas de privacidade diferencial e NER, atingindo F1-score de 0.576 com o modelo KeyBERT. Esses estudos evidenciam a crescente preocupação com a privacidade em diferentes idiomas, domínios e modalidades (texto e fala), demonstrando que, embora haja avanços técnicos, ainda há desafios relacionados à adaptação linguística, recursos limitados e equilíbrio entre anonimização e preservação da informação.

A partir da análise realizada pode-se relacionar desafios comuns entre os artigos selecionados e o presente estudo. Um desafio significativo está relacionado à escassez de dados para o treinamento dos modelos, questão recorrente devido a sensibilidade imposta pela própria natureza dos dados a serem reconhecidos. É observado que os corpora utilizados em sua maioria são baseados em documentos e textos [Zhang et al. 2023] e [Catelli et al. 2021]. Além disso, podemos observar que todos os estudos realizam as suas análises considerando a métrica F1 (F1-score), que combina as métricas de precisão e *recall*. Outro ponto comum entre os estudos foi a utilização dos modelos BERT e Bi-LSTM+CRF, empregados por [Catelli et al. 2020], [Gultiaev and Domashova 2022], [Herwanto et al. 2021], [Hu et al. 2022], [Wongvises et al. 2022] e [Zhang et al. 2023].

O modelo que será empregado neste trabalho será o BERT, escolhido com base nos trabalhos revisados. Como principal diferença, o presente estudo propõe o reconhecimento de entidades nomeadas para identificação de dados pessoais em transcrições de áudios na língua portuguesa. Este estudo possui enfoque na manutenção da privacidade do titular, utilizando-se também da capacidade de geração de dados sintéticos por meio de modelos de linguagem de grande escala (*Large Language Model* - LLM).

3. Experimento

A estrutura do experimento proposto para este estudo é delineada em três etapas principais, conforme ilustrado na Figura 1, que descreve o fluxo processual e a interconexão entre as fases do estudo: Geração de Dados Sintéticos (Corpus), Pré-Processamento e Extração de Informações. Cada etapa é essencial para o desenvolvimento e a avaliação dos modelos de NER aplicados a transcrições de áudio, com o objetivo de reconhecer e classificar entidades de dados pessoais. Todos os códigos desenvolvidos estão disponíveis em repositório público no GitHub[1].

3.1. Estabelecimento do Corpus

Como discutido por [Nikolenko 2019], o uso de dados sintéticos é motivado pela necessidade de contornar restrições legais e garantir privacidade. Este trabalho trata da descoberta de dados pessoais, justificando assim o uso de dados artificiais. Foram consideradas as seguintes entidades: nome, CPF, RG e endereço. Além disso, foram anotadas duas relações: ”Reside em” e ”Possui Documento”.

[1]https://github.com/mskca/tcc-nlp-2024

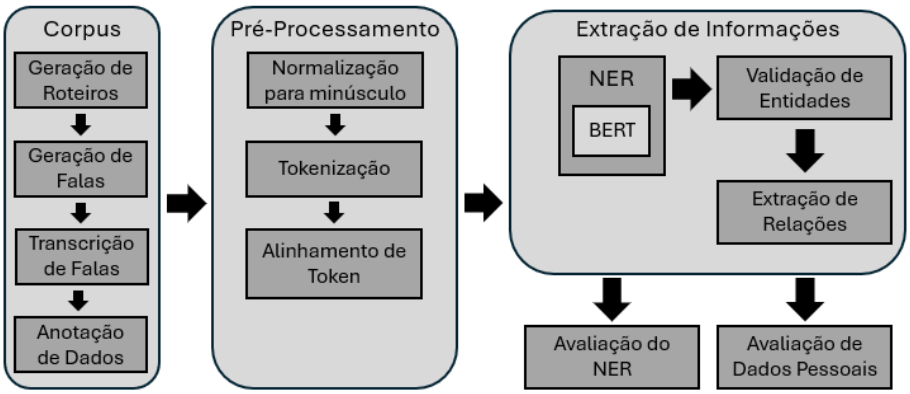

Figura 1. Etapas do Processo de NLP

A primeira fase do experimento envolve a criação do corpus, constituído por roteiros gerados e suas falas correspondentes. A geração dos dados sintéticos foi feita com a plataforma 4devs[2], utilizada especificamente para CPF, RG e endereço. A escolha da ferramenta ocorreu pela sua capacidade de fornecer dados variados e realistas, importantes para um corpus representativo. Os RGs seguem o padrão da SSP-SP. Os nomes foram obtidos a partir do censo oficial do IBGE[3]. Todos os endereços foram padronizados para o estado do Rio Grande do Sul.

Definiu-se um cenário realista baseado em diálogos entre um atendente de suporte técnico e um cliente. Os roteiros foram gerados com o modelo GPT-4o[4], utilizando os dados pessoais criados anteriormente. Cada roteiro foi estruturado em arquivos *JSON*, como ilustrado na Figura 2.

```
{
    "falas": [
        {
            "nome": "Izabel",
            "genero": "F",
            "tipo": "atendente",
            "fala": "Boa tarde, meu nome é Izabel. Com quem eu falo, por favor?"
        },
        {
            "nome": "Patricia",
            "genero": "F",
            "tipo": "usuario",
            "fala": "Boa tarde, Izabel. Aqui é a Patricia."
        },
        {
            "nome": "Izabel",
            "genero": "F",
            "tipo": "atendente",
            "fala": "Oi, Patricia. Em que posso ajudá-la hoje?"
        },
        {
            "nome": "Patricia",
            "genero": "F",
            "tipo": "usuario",
            "fala": "Eu estou com um problema referente ao meu CPF 698.237.760-16."
        },
```

Figura 2. Exemplo de Roteiro

Em seguida, os roteiros foram transformados em áudios com o modelo ele-

[2]https://www.4devs.com.br/
[3]https://www.ibge.gov.br/
[4]https://platform.openai.com/docs/api-reference

ven_multilingual_v2[5], permitindo gerar diálogos com entonações e sotaques variados. As transcrições foram feitas com o modelo da AssemblyAI[6], garantindo boa precisão na conversão. As transcrições foram então anotadas com a ferramenta LabelBox[7], onde foram rotuladas entidades e suas relações. A Figura 3 mostra um exemplo de transcrição anotada.

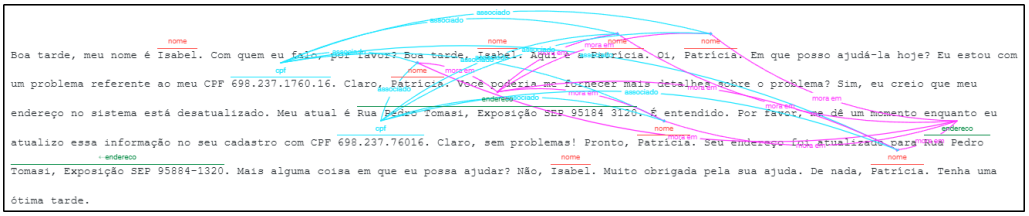

Figura 3. Exemplo de Transcrição Anotada.

Para garantir diversidade e representatividade, foram criados 500 roteiros com duração média de 57 segundos, totalizando aproximadamente 8 horas de áudio. As anotações geraram 3.354 Nomes, 328 CPFs, 306 RGs, 539 Endereços, além de 2.185 casos de ”Reside em”e 2.607 de ”Possui Documento”. Cada roteiro contém, em média, até um CPF ou RG, até dois nomes e um endereço, assegurando uma boa distribuição das entidades.

3.2. Pré-Processamento

Neste estudo, o pré-processamento foi aplicado no arquivo anotado em formato ndjson e consistiu em três principais tarefas: normalização para minúsculo, tokenização e alinhamento de tokens. A normalização converte todo o texto para letras minúsculas, eliminando variações derivadas da capitalização e promovendo uniformidade no conjunto de dados [Campesato 2020]. Por exemplo, as palavras ”Texto”e ”texto” passam a ser tratadas da mesma forma após a normalização. Já a tokenização divide o texto em unidades menores chamadas tokens, que podem ser palavras, subpalavras ou caracteres, permitindo ao modelo processar o conteúdo de forma mais granular. O modelo BERT, utilizado neste trabalho, adota uma tokenização baseada em subpalavras, o que melhora sua capacidade de lidar com variações morfológicas e semânticas [Eisenstein 2019]. Para essa tarefa, foi utilizada a biblioteca *transformers* da Hugging Face[8], que fornece ferramentas eficientes e adaptadas ao uso com BERT.

A terceira etapa do pré-processamento é o alinhamento de tokens, essencial para tarefas de rotulagem sequencial, como o reconhecimento de entidades mencionadas. Esse processo garante que cada token esteja corretamente associado à posição correspondente no texto original. Um método comum de alinhamento envolve o uso de *offsets*, que registram a posição exata de cada token no texto, permitindo rastreamento preciso. Adicionalmente, são utilizadas as etiquetas do esquema BIO (*Begin*, *Inside* e *Outside*) para indicar a função de cada token no contexto das entidades, o que torna possível representar de forma estruturada e padronizada a presença e os limites das entidades no texto. Esse conjunto

[5]https://elevenlabs.io/docs/introduction
[6]https://www.assemblyai.com/docs/
[7]https://labelbox.com/
[8]https://huggingface.co/docs/transformers/en/index

de técnicas assegura maior consistência e qualidade nos dados de entrada, impactando diretamente o desempenho dos modelos treinados.

3.3. Extração de Informações

Esta seção descreve as técnicas de extração de informações utilizadas para identificar dados pessoais em transcrições de áudio, focando nas etapas de reconhecimento de entidades nomeadas (NER) e extração de relações entre essas entidades.

O reconhecimento de entidades nomeadas foi realizado utilizando o modelo BERT, escolhido por sua eficácia em tarefas no idioma Português [Ignaczak et al. 2023]. Essa etapa é essencial para identificar e classificar corretamente as entidades sensíveis como "Nome", "CPF", "RG" e "Endereço". Para isso, foi utilizado o modelo pré-treinado *bert-base-portuguese-cased*, disponível na plataforma Hugging Face[9], posteriormente adaptado por meio de *fine-tuning* com um corpus específico. A adaptação seguiu uma abordagem supervisionada de classificação multiclasse, utilizando dados em formato ndjson.

O corpus foi dividido em 80% para treinamento e 20% para teste, e o modelo foi treinado por três épocas utilizando a técnica de *K-Fold Cross Validation* com $k = 10$. Essa abordagem permitiu mitigar a limitação do pequeno conjunto de dados rotulados, pois cada subconjunto atua alternadamente como teste e treino, contribuindo para uma avaliação mais robusta do desempenho médio dos modelos gerados [Eisenstein 2019]. Durante a validação, além da avaliação por métricas quantitativas, houve atenção especial à precisão contextual das entidades extraídas, assegurando que o modelo reconhecesse não apenas as entidades com acurácia, mas também sua relevância no contexto da transcrição.

Após o reconhecimento das entidades, foi aplicada uma abordagem de extração de relações, na qual o modelo analisa a proximidade e o contexto das entidades em uma sentença para inferir vínculos semânticos. Neste estudo, foram definidas duas relações principais: "Reside em" e "Possui documento". A relação "Reside em" associa a entidade "Nome" à entidade "Endereço" informada pelo mesmo cliente na transcrição, enquanto a relação "Possui documento" vincula o "Nome" às entidades "CPF" ou "RG" extraídas do texto. Essas relações refletem estruturas comuns em chamadas de atendimento e são relevantes para fins de conformidade com a LGPD [Eisenstein 2019].

O treinamento do modelo de extração de relações seguiu a mesma configuração do NER: três épocas, validação cruzada com $k = 10$ e divisão de 80/20 para o treinamento e teste. Essa consistência metodológica visa garantir comparabilidade entre os modelos e maximizar o aprendizado mesmo com um corpus limitado. A identificação automática dessas relações pode fortalecer sistemas de monitoramento de conformidade e auditoria de dados, permitindo que organizações compreendam melhor como informações sensíveis são mencionadas e vinculadas em interações verbais.

4. Resultados

Nesta seção são apresentados os resultados obtidos a partir do experimento realizado com o modelo BERTimbau para o reconhecimento de entidades nomeadas e extração

[9]https://huggingface.co/neuralmind/bert-base-portuguese-cased

de relações. As Seções 4.1, 4.2 e 4.3 detalham respectivamente os resultados do NER, da extração de relações e a avaliação geral do experimento realizado.

A Tabela 2 exibe o resultado geral do desempenho do modelo. Observa-se que o BERTimbau apresentou desempenho significativo no NER, com uma média de F1-score de 0.98. Em contraste, o desempenho da extração de relações foi significativamente inferior, com um F1-score de 0.29, destacando a dificuldade do modelo em identificar corretamente as relações. Já a Tabela 3 detalha os resultados por entidade e por relação, destacando o desempenho do modelo em cada categoria específica. Para o reconhecimento de entidades, o modelo BERTimbau obteve alta precisão e *recall* em todas as classes. No entanto, os resultados para a extração de relações demonstram um desempenho consideravelmente inferior.

	Precisão	**Recall**	**F1-score**
NER	0.97	0.99	0.98
Extração de Relações	0.23	0.45	0.29

Tabela 2. Performance geral do modelo refinado para reconhecimento de entidades e relações.

	Classificação	**Métricas**		
		Precisão	**Recall**	**F1-score**
NER (BERTimbau)	Nome	0.98	1.00	0.99
	CPF	0.96	0.98	0.97
	RG	0.97	0.99	0.98
	Endereço	0.98	1.00	0.99
Extração de Relações	Reside em	0.44	0.95	0.60
	Possui documento	0.06	0.06	0.06

Tabela 3. Performance do modelo refinado por entidade e relação.

4.1. Reconhecimento de Entidades Nomeadas

A Figura 4 apresenta dois exemplos de entidades válidas detectadas: ”CPF” e ”Endereço”. A segmentação dos tokens segue a metodologia ”BIO”, permitindo a reconstrução das entidades a partir dos rótulos atribuídos. Mesmo em casos em que o padrão textual não corresponde a um formato típico, como um CPF incorreto, o modelo foi capaz de identificar corretamente a entidade com base no contexto, caracterizando um verdadeiro positivo.

```
CPF Detectado: 28 ##7 .99 ##6 . 39 ##0 .
Endereço Detectado: rua fonte nova são tom ##é ce ##p 94 ##46 ##60 04
```

Figura 4. Exemplo de detecção

De acordo com a Tabela 3 o modelo BERTimbau demonstrou desempenho elevado no reconhecimento das quatro entidades. Para ”Nome”e ”Endereço”, o modelo atingiu F1-score de 0.99, com *recall* de 1.00, indicando que todas as ocorrências dessas entidades foram corretamente identificadas. ”CPF”e ”RG”também obtiveram altos resultados, com F1-scores de 0.97 e 0.98, respectivamente. Esses números refletem a capacidade

do modelo de generalizar padrões mesmo com variações textuais mínimas. No geral, o desempenho obtido reforça o potencial do BERTimbau para aplicações em conformidade com regulamentações de privacidade, como a LGPD.

4.2. Extração de Relações

A extração de relações apresentou desempenho inferior quando comparada à tarefa de NER. A relação ”Reside em”, que conecta a entidade ”Nome” à entidade ”Endereço”, obteve um F1-score de 0.60, com *recall* elevado (0.95) e precisão consideravelmente baixa (0.44). Esses resultados indicam que, embora o modelo tenha sido capaz de identificar a maioria dos casos em que a relação realmente existia, também incorreu em um grande número de falsos positivos — inferindo relações inexistentes. Esse comportamento pode estar relacionado à padronização presente nos dados sintéticos, que repetem estruturas como “meu nome é... e moro em...”, levando o modelo a generalizar excessivamente. A ausência de variação e a previsibilidade nas construções sintáticas contribuem para que o modelo encontre padrões onde não há, prejudicando a precisão.

Por outro lado, a relação ”Possui documento”, que liga ”Nome”a ”CPF”ou ”RG”, apresentou desempenho crítico, com F1-score, precisão e *recall* todos fixados em 0.06. Isso sugere falha do modelo em identificar corretamente esse tipo de vínculo semântico. A baixa performance pode ser explicada pela escassez de exemplos no corpus sintético, além da complexidade inerente à interpretação contextual — a relação só é válida quando o ”Nome” refere-se ao cliente, e não ao atendente, o que requer um nível de compreensão mais profundo do diálogo. No geral, a tarefa de extração de relações obteve F1-score médio de 0.29, com *recall* de 0.45 e precisão de apenas 0.23, indicando uma tendência do modelo a reconhecer muitas relações, mas com grande número de falsos positivos. Melhorias nesse aspecto podem ser alcançadas com a introdução de dados mais realistas e variados, bem como com o uso de técnicas mais refinadas de ajuste fino e *context-awareness*, capazes de aprimorar a capacidade de generalização e reduzir erros de inferência.

4.3. Discussão dos Resultados

Os resultados obtidos demonstram que o modelo BERTimbau apresenta desempenho eficaz na tarefa de NER ao lidar com informações pessoais como nome, CPF, RG e endereço. O alto F1-score médio de 0.98 evidencia o potencial da aplicação desse modelo em cenários de automação, verificação de identidade e conformidade com legislações como a LGPD. No entanto, é importante reconhecer que parte desse desempenho pode estar associado ao uso de um corpus sintético, cuja estrutura previsível facilita o aprendizado do modelo e pode inflar artificialmente os resultados. A padronização das expressões nos dados gerados tende a criar um ambiente mais favorável para o reconhecimento, reduzindo a variabilidade linguística encontrada em contextos reais.

A análise do corpus revela um viés gerado pela repetição de padrões como “O meu CPF é...” ou “O meu endereço é...”, o que pode ter influenciado o modelo a memorizar estruturas específicas em vez de aprender padrões linguísticos mais amplos. Embora esse fator tenha contribuído para a eficácia do modelo em NER, ele também limita sua capacidade de generalização. Portanto, torna-se essencial o desenvolvimento de dados sintéticos mais diversos e realistas, com maior variação semântica e sintática, a fim de mitigar esse viés e promover uma aprendizagem mais robusta. Em contrapartida, a extração de relações demonstrou desempenho significativamente inferior, com F1-score médio de

0.29, especialmente em relações como "Possui documento". Isso reflete a dificuldade do modelo em compreender relações semânticas complexas com um conjunto de dados limitado e com pouca diversidade expressiva, sugerindo a necessidade de aprimoramento no corpus e em técnicas específicas para essa tarefa.

Apesar das limitações observadas, o uso de dados sintéticos ainda se mostra uma solução prática e estratégica para o desenvolvimento de modelos de PLN, especialmente em domínios sensíveis onde o acesso a dados reais é restrito. O sucesso na tarefa de NER reforça a viabilidade dessa abordagem, mesmo considerando as limitações em relação à extração de relações. Além disso, embora não seja possível realizar uma comparação diretamente com os resultados obtidos em outros estudos, observa-se que o resultado da métrica F1-score obtida neste estudo supera outros modelos BERT, como [Herwanto et al. 2021, Hu et al. 2022, Wongvises et al. 2022], cujos valores ficaram abaixo de 0.73. No grupo de estudos voltados à privacidade, aproximadamente 33% atingiram F1-score acima de 0.91, o que posiciona os resultados deste trabalho de forma bastante positiva, principalmente considerando o uso exclusivo de dados sintéticos e a restrição de recursos computacionais.

5. Considerações Finais

Este trabalho investigou o uso de técnicas de reconhecimento de entidades nomeadas (NER) para identificar dados pessoais — como nome, CPF, RG e endereço — em transcrições de áudio em português. Para isso, foi desenvolvido e avaliado um modelo baseado no BERTimbau, utilizando dados sintéticos como alternativa à escassez de dados reais anotados. A motivação central é a crescente demanda por soluções automatizadas que atendam à LGPD, possibilitando o uso seguro e eficiente de grandes volumes de dados sensíveis em ambientes corporativos e institucionais.

Os resultados mostraram que o modelo teve desempenho excelente em NER, com F1-score médio de 0,98, comprovando sua eficácia na identificação de entidades pessoais. Esse desempenho reforça seu potencial em sistemas automatizados de verificação de identidade e auditoria. Por outro lado, a extração de relações obteve F1-score médio de apenas 0,29, refletindo a maior complexidade da detecção de vínculos semânticos. Além disso, o uso de dados sintéticos, embora útil para NER, gerou padrões repetitivos que podem ter introduzido viés e reduzido a generalização do modelo em cenários reais, mais variados e imprevisíveis.

Os resultados obtidos neste estudo contribuem para demonstrar a viabilidade da descoberta de entidades associadas a dados pessoais em transcrições de áudios em Português. Além disso, apesar da menção à possibilidade de introdução de viés no modelo avaliado, o trabalho apresenta a viabilidade do uso de corpus sintético para o treinamento de modelos com o objetivo de identificar dados pessoais, pois pesquisadores não poderão contar com corpora públicos com este tipo de dados devido às questões éticas e legais.

Entre as limitações deste estudo estão o uso de apenas um modelo e a falta de comparações com outras arquiteturas. Futuramente, recomenda-se o uso de dados reais anotados fornecidos por organizações de diferentes domínios e a avaliação de outras abordagens para NER e extração de relações. A geração de dados sintéticos mais diversos também deve ser explorada para mitigar vieses e ampliar a robustez dos modelos. Essas medidas podem resultar em soluções mais completas e generalizáveis.

Referências

ABNT (2023). ABNT NBR ISO/IEC 27005:2023 - Tecnologia da Informação – Técnicas de Segurança – Gestão de riscos de segurança da informação. International Organization for Standardization. Acesso em: 20 maio 2024.

Aggarwal, C. C. and Zhai, C., editors (2012). *Mining Text Data*. Springer, New York, NY, 1 edition.

Bannour, N., Wajsbürt, P., Rance, B., Tannier, X., and Névéol, A. (2022). Privacy-preserving mimic models for clinical named entity recognition in french. *Journal of Biomedical Informatics*, 130:104073.

Brasil (2018). Lei nº 13.709, de 14 de agosto de 2018. *Diário Oficial [da] República Federativa do Brasil*.

Campesato, O. (2020). *Artificial Intelligence, Machine Learning, and Deep Learning*. Mercury Learning and Information, Berlin, Boston.

Catelli, R., Gargiulo, F., Casola, V., De Pietro, G., Fujita, H., and Esposito, M. (2020). Crosslingual named entity recognition for clinical de-identification applied to a covid-19 italian data set. *Applied Soft Computing*, 97:106779.

Catelli, R., Gargiulo, F., Damiano, E., Esposito, M., and De Pietro, G. (2021). Clinical de-identification using sub-document analysis and electra. In *2021 IEEE International Conference on Digital Health (ICDH)*, pages 266–275.

Eisenstein, J. (2019). *Introduction to natural language processing*. The MIT Press.

Gartner (2019). Gartner predicts 2019 for the future of privacy. Acesso em: 17 nov. 2024.

Gultiaev, A. A. and Domashova, J. V. (2022). Developing a named entity recognition model for text documents in russian to detect personal data using machine learning methods. *Procedia Computer Science*, 213:127–135. 2022 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence: The 13th Annual Meeting of the BICA Society.

Herwanto, G. B., Quirchmayr, G., and Tjoa, A. M. (2021). A named entity recognition based approach for privacy requirements engineering. In *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, pages 406–411.

Hu, Y., Li, R., Wang, S., Tao, F., and Sun, Z. (2022). Speechhide: A hybrid privacy-preserving mechanism for speech content and voiceprint in speech data sharing. In *2022 7th IEEE International Conference on Data Science in Cyberspace (DSC)*, pages 345–352.

Ignaczak, L., Martins, M. G., da Costa, C. A., Donida, B., and da Silva, M. C. P. (2023). An evaluation of nerc learning-based approaches to discover personal data in brazilian portuguese documents. *Discover Data*, 1(1).

Moussaoui, T. E., Chakir, L., and Boumhidi, J. (2023). Preserving privacy in arabic judgments: Ai-powered anonymization for enhanced legal data privacy. *IEEE Access*, 11:117851–117864.

Neves, M. (2022). O que são dados e por que eles são importantes? Nubank Blog. Acesso em: 20 maio 2024.

Nikolenko, S. I. (2019). Synthetic data for deep learning. *arXiv preprint arXiv:1909.11512*.

Silva, P., Gonçalves, C., Godinho, C., Antunes, N., and Curado, M. (2020). Using nlp and machine learning to detect data privacy violations. In *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 972–977.

Wongvises, C., Khurat, A., and Noraset, T. (2022). Thai privacy notice analysis based on named-entity recognition technique. In *2022 26th International Computer Science and Engineering Conference (ICSEC)*, pages 257–262.

Zhang, B., Yao, X., Li, H., and Aini, M. (2023). Chinese medical named entity recognition based on expert knowledge and fine-tuning bert. In *2023 IEEE International Conference on Knowledge Graph (ICKG)*, page 84–90. IEEE.