CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

RECONHECIMENTO DE ENTIDADES NOMEADAS EM RELATOS CRIMINAIS INFORMAIS COM APOIO DE METADADOS ESTRUTURADOS

Gustavo Melo

Proposta de dissertação apresentada ao Programa de Pós-graduação em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ, como parte dos requisitos necessários à obtenção do grau de mestre.

Orientadores:
Eduardo Bezerra, D.Sc.
Karla Figueiredo, D.Sc.

Rio de Janeiro,

1 de Julho de 2026

Reconhecimento de Entidades Nomeadas em Relatos Criminais Informais com Apoio de Metadados Estruturados

Proposta de dissertação em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ.

Gustavo Melo

Aprovada por:

Presidente, Eduardo Bezerra, D.Sc. (CEFET/RJ)

Karla Figueiredo, D.Sc. (UERJ)

Ronaldo Ribeiro Goldschmidt, D.Sc. (IME/RJ)

Gustavo Paiva Guedes, D.Sc. (CEFET/RJ)

Kele Teixeira Belloze, D.Sc. (CEFET/RJ)

Rio de Janeiro,

1 de Julho de 2026

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

M528 Melo, Gustavo
Reconhecimento de entidades nomeadas em relatos criminais informais com apoio de metadados estruturados / Gustavo Melo. — 2026.
100f. + apêndice : il. color. , enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2026.
Bibliografia : f. 96-100
Orientador: Eduardo Bezerra
Coorientadora: Karla Figueiredo

1. Processamento de linguagem natural (Computação). 2. Segurança pública. 3. Metadados. 4. Segurança pública – Participação do cidadão. I. Bezerra, Eduardo. (Orient.). II. Figueiredo, Karla. (Coorient.). III. Título.

CDD 006.35

Elaborada pela bibliotecária Tania Mello – CRB/7 nº 5507/04

As coisas estão no mundo

Só que eu preciso aprender

(Paulo César Batista de Faria)

Agradecimentos

Agradeço aos meus pais, Ciléa e João Roberto, pelo apoio, incentivo e determinação em tornar possível a realização do meu mestrado.

Agradeço aos meus amigos pela amizade e pelos ombros amigos nos momentos mais difíceis, e à minha companheira, Paula, pelo apoio e pela ajuda em me manter focado nos estudos.

Agradeço às músicas e aos livros que fazem parte da minha vida, trazendo leveza e novas visões sobre o que o mundo pode ser.

Agradeço à minha coorientadora, Karla Figueiredo, pelos ensinamentos desde a graduação e pela oportunidade de trazer esses estudos para o CEFET-RJ, ampliando as pesquisas já realizadas.

Agradeço ao meu orientador, Eduardo Bezerra, pela paciência na orientação e pelo apoio fundamental nos momentos mais difíceis.

Agradeço também ao Instituto de Segurança Pública e ao Disque Denúncia pela colaboração, parceria e troca de informações, que foram indispensáveis para a concretização deste trabalho.

Por fim, gostaria de agradecer a todo o corpo docente do PPCIC/CEFET-RJ pela transmissão do conhecimento essencial para a formação da minha carreira.

RESUMO

Reconhecimento de Entidades Nomeadas em Relatos Criminais Informais com Apoio de Metadados Estruturados

Gustavo Melo

Orientadores:
Eduardo Bezerra, D.Sc.
Karla Figueiredo, D.Sc.

Resumo da Dissertação submetida ao Programa de Pós-graduação em Ciência da Computação do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ como parte dos requisitos necessários à obtenção do grau de mestre.

O presente trabalho investiga o problema do reconhecimento de entidades nomeadas em relatos criminais informais registrados no serviço Disque Denúncia. Esses relatos, frequentemente marcados por linguagem coloquial, erros ortográficos e estrutura textual livre, impõem desafios significativos ao uso de modelos tradicionais de Processamento de Linguagem Natural. Além do texto livre, os relatos são acompanhados de metadados estruturados, como tipo de ocorrência, localização e data que podem fornecer contexto adicional relevante para a tarefa. Neste estudo, propõe-se uma abordagem baseada em *fine-tuning* de modelos de linguagem de grande porte, utilizando um corpus*m*anualmente anotado, por quatro diferentes pessoas e pelo autor da dissertação, com entidades do tipo *Pessoa*, *Local* e *Organização*. Para contornar a escassez de dados rotulados, a metodologia inclui a aplicação de pseudo-rotulagem em um segundo *corpus*, significativamente maior, expandindo assim a base de treinamento de forma semissupervisionada. Adicionalmente, os metadados dos relatos são incorporados como fonte de contexto tanto no pré-processamento quanto no processo de avaliação e refinamento dos modelos. Os experimentos foram conduzidos com o modelo GliNER e avalia que a utilização de metadados e pseudo-rotulação pode contribuir para melhorar a performance dos modelos em *corpora* informais, com impactos positivos para a extração automatizada de informações em contextos de segurança pública. Os resultados reforçam o potencial de abordagens híbridas e sensíveis ao domínio para aplicações reais de Processamento de Linguagem Natural em ambientes com escassez de rótulos anotados.

Palavras-chave:
Reconhecimento de Entidades Nomeadas, Processamento de Linguagem Natural, Linguagem Informal, Segurança Pública, Fine-Tuning, Pseudo-Rotulagem, Metadados Estruturados

Rio de Janeiro,

1 de Julho de 2026

ABSTRACT

Reconhecimento de Entidades Nomeadas em Relatos Criminais Informais com Apoio de Metadados Estruturados

Gustavo Melo

Advisors:
Eduardo Bezerra, D.Sc.
Karla Figueiredo, D.Sc.

Abstract of dissertation submitted to Programa de Pós-graduação em Ciência da Computação - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ as partial fulfillment of the requirements for the degree of master.

This work investigates the problem of Named Entity Recognition (NER) in informal criminal reports registered through the Disque Denúncia service. These reports, often characterized by colloquial language, spelling errors, and free-text structure, pose significant challenges to the application of traditional Natural Language Processing models. In addition to free text, the reports are accompanied by structured metadata, such as occurrence type, location, and date, which may provide additional contextual information relevant to the task. In this study, an approach based on *fine-tuning* large language models is proposed, using a manually annotated *corpus*, annotated by four different individuals and the dissertation author, containing entities of type *Person*, *Location*, and *Organization*. To address the scarcity of labeled data, the methodology includes the application of pseudo-labeling on a second and significantly larger *corpus*, thus expanding the training set in a semi-supervised manner. Additionally, report metadata are incorporated as a contextual source both during preprocessing and throughout the evaluation and refinement stages of the models. The experiments were conducted using the GliNER model and evaluate how the use of metadata and pseudo-labeling can contribute to improving model performance in informal *corpora*, with positive impacts on automated information extraction in public security contexts. The results reinforce the potential of hybrid and domain-sensitive approaches for real-world Natural Language Processing applications in environments with limited annotated data.

Keywords:
Named Entity Recognition, Natural Language Processing, Informal Language, Public Safety, Fine-Tuning, Pseudo-Labeling, Structured Metadata

Rio de Janeiro,

1 de Julho de 2026

Sumário

I Introdução **1**

I.1 Contextualização 1

I.2 Motivação 3

I.3 Objetivos 4

I.4 Visão Geral da Abordagem 4

I.5 Organização 5

II Referencial Teórico **7**

II.1 Reconhecimento de Entidades Nomeadas (REN) 7

II.1.1 Definição e Histórico 7

II.1.2 Esquemas de anotação 8

II.1.3 Métricas de avaliação 10

II.2 Transformers e Modelos Pré-treinados 13

II.2.1 Transformers 13

II.2.2 Modelos de Linguagens Pré-Treinados 14

II.2.3 GliNER 15

II.2.4 Ajuste Fino 19

II.3 Outras Técnicas Usadas neste Trabalho 20

II.3.1 Validação Cruzada 20

II.3.2 Pseudo-Rotulagem 21

II.3.3 Calibração de Modelos 23

II.4 Discussão 26

III Trabalhos Relacionados **27**

III.1 Seleção de publicações 27

III.2 Critérios de inclusão, exclusão e priorização 29

III.2.1 Critérios de Inclusão 30

III.2.2 Critérios de Exclusão 30

III.2.3 Critérios de Priorização 31

III.3 Resultados 31
III.3.1 Publicações selecionadas 31
III.3.2 Comparação 34
III.4 Discussão 37

IV Reconhecimento de Entidades Nomeadas em Relatos Criminais 38
IV.1 Formulação do Problema 38
IV.2 *Corpora* de Relatos Criminais 39
IV.3 Etapas da metodologia 45
IV.3.1 Pré-Processamento dos Relatos Criminais 46
IV.3.2 Anotação Manual do corpus DD-small 47
IV.3.3 Divisão do corpus DD-small 49
IV.3.4 Treinamento do Modelo de Referência 50
IV.3.5 Aplicação de Auto-treino Iterativo 50
IV.3.6 Avaliação 59

V Experimentos Computacionais 62
V.1 Plano Experimental 62
V.2 Análise Exploratória 64
V.2.1 Rotulação dos Relatos Criminais da Validação Manual 64
V.2.2 Rotulação dos Relatos Criminais da Pseudo-rotulação 65
V.3 Configurações Experimentais 67
V.4 Comparando alternativas do modelo GliNER 68
V.5 Comparação com outros modelos 69
V.5.1 Prompts Utilizados nos Modelos GPT e Gemini 71
V.6 Calibração da pseudo-rotulação do corpus DD-large 72
V.6.1 Métricas da Calibração 74
V.7 Impacto da pseudo-rotulação do corpus DD-large 77
V.7.1 Utilização de Metadados 78
V.7.2 Sem Utilizar os metadados 81
V.8 Discussão 88

VI Conclusão 93
VI.1 Análise Retrospectiva 93
VI.2 Trabalhos Futuros 94

Referências 96

Lista de Figuras

II.1 Arquitetura do modelo *Transformer*. 14

II.2 Arquitetura do modelo GliNER. 17

II.3 Esquema de pseudo-rotulação 22

III.1 Fluxo de seleção de artigos segundo a metodologia PRISMA. 29

IV.1 Proporção de palavras por relato – DD_{small} e DD_{large} 42

IV.2 Fluxograma da metodologia proposta. 46

IV.3 Fluxograma do treinamento do modelo original M_0 sobre o *corpus* DD_{small}. 51

IV.4 Fluxo do auto-treinamento iterativo 52

V.1 Frequência da confiança das entidades – DD_{large} 73

V.2 Frequência da confiança das entidades – DD_{large} com o Modelo *Temperature Scaling* 73

V.3 Frequência da confiança das entidades – DD_{large} com o Modelo *Isotonic Regression* 74

V.4 Curva de Confiabilidade dos modelos 75

V.5 Matriz de confusão do modelo sem calibração 79

V.6 Matriz de confusão do modelo com calibração por *Temperature Scaling* 80

V.7 Matriz de confusão do modelo com calibração por *Isotonic Regression* 82

V.8 Matriz de confusão sem uso de metadados e sem calibração 84

V.9 Matriz de confusão sem metadados com calibração por *Temperature Scaling* 86

V.10 Matriz de confusão sem metadados com calibração por *Isotonic Regression* 88

A.1 Nuvem de palavras gerada a partir das entidades extraídas dos relatos criminais no corpus $DD_{\mathbf{small}}$. As entidades mais frequentes são representadas com maior destaque por meio do aumento proporcional do tamanho da fonte. 101

A.2 Nuvem de palavras gerada a partir das entidades de Localização extraídas dos relatos criminais no corpus $DD_{\mathbf{small}}$. As entidades mais frequentes são representadas com maior destaque por meio do aumento proporcional do tamanho da fonte. 102

A.3 Nuvem de palavras gerada a partir das entidades de Pessoa extraídas dos relatos criminais no corpus $DD_{\mathbf{small}}$. As entidades mais frequentes são representadas com maior destaque por meio do aumento proporcional do tamanho da fonte. 102

A.4 Nuvem de palavras gerada a partir das entidades de Pessoa extraídas dos relatos criminais no corpus $DD_{\mathbf{small}}$. As entidades mais frequentes são representadas com maior destaque por meio do aumento proporcional do tamanho da fonte. 103

A.5 Nuvem de palavras gerada a partir das entidades extraídas dos relatos criminais no corpus $DD_{\mathbf{large}}$. As entidades mais frequentes são representadas com maior destaque por meio do aumento proporcional do tamanho da fonte. 103

A.6 Nuvem de palavras gerada a partir das entidades de Localização extraídas dos relatos criminais no corpus $DD_{\mathbf{large}}$. As entidades mais frequentes são representadas com maior destaque por meio do aumento proporcional do tamanho da fonte. 104

A.7 Nuvem de palavras gerada a partir das entidades de Pessoa extraídas dos relatos criminais no corpus $DD_{\mathbf{large}}$. As entidades mais frequentes são representadas com maior destaque por meio do aumento proporcional do tamanho da fonte. 104

A.8 Nuvem de palavras gerada a partir das entidades de Organização extraídas dos relatos criminais no corpus $DD_{\mathbf{large}}$. As entidades mais frequentes são representadas com maior destaque por meio do aumento proporcional do tamanho da fonte. 105

Lista de Tabelas

II.1 Comparação entre abordagens de avaliação em REN 12

III.1 Visão geral das características dos trabalhos relacionados. 36

IV.1 Resumo dos *corpora* 40

IV.2 Campos de metadados existentes em cada relato. 41

IV.3 Proporção por categoria no DD_{small} e DD_{large} 41

IV.4 Estatísticas gerais dos corpora 43

IV.5 Completude e Cardinalidade do $DD_{\mathbf{large}}$. 43

IV.6 Dez valores mais recorrentes dos Metadados do $DD_{\mathbf{large}}$ 44

IV.7 Resumo Estatístico sobre entidades nomeadas 45

IV.8 Exemplos do formato BIO 47

IV.9 Frequência do *corpus* da validação manual 48

IV.10 Distribuição das entidades 48

IV.11 Métricas de entidades e frequência no DD_{small} 49

IV.12 Frequência de Entidades nos Conjuntos de Treino e Teste 49

V.1 Exemplos BIO para entidades de local 64

V.2 Exemplos BIO para entidades de pessoa 65

V.3 Exemplos BIO para entidades de organização 65

V.4 Exemplos BIO de entidades de local após ajuste fino 66

V.5 Exemplos BIO de entidades de pessoa após ajuste fino 66

V.6 Exemplos BIO de entidades de organização após ajuste fino 67

V.7 Resultados nos Testes GliNER. 69

V.8 Comparação entre os modelos testados 70

V.9 Resultados das métricas da calibração 76

V.10 Resultados obtidos nas iterações de pseudo-rotulação sem calibração dos modelos. 78

V.11 Resultados obtidos nas iterações de pseudo-rotulação utilizando *Temperature Scaling*. 80

V.12 Resultados obtidos nas iterações de pseudo-rotulação utilizando o modelo *Isotonic Regression* para a calibração dos modelos e utilizando os metadados. 81

V.13 Resultados das iterações de pseudo-rotulação sem uso de metadados e sem calibração. 83
V.14 Resultados das iterações de pseudo-rotulação sem metadados e com calibração por *Temperature Scaling*. 85
V.15 Resultados das iterações de pseudo-rotulação sem metadados e com calibração por *Isotonic Regression*. 87

Lista de Abreviações

ACL	*Association for Computational Linguistics*	27, 29, 37
BERT	*Bidirectional Encoder Representations from Transformers*	7, 14, 15, 16, 18, 33, 34, 35, 37, 89
BILOU	*Begin, Inside, Last, Outside, Unit*	8, 9, 15
BI-LSTM	*Bidirectional recurrent neural networks*	33
BID	Banco Interamericano De Desenvolvimento	1
BILSTM	*Bidirectional Long Short-Term Memory*	35
BIO	*Begin-Inside-Outside*	8, 9, 15, 39, 47, 64, 65, 71
CRF	*Conditional Random Field*	35
DEBERTA	*Decoding-enhanced BERT with Disentangled Attention*	15
ECE	*Expected Calibration Error*	57, 58, 76
EPD	*Euphemistic Phrase Detection*	33
GLINER	*Generalist and Lightweight Model for Named Entity Recognition using Bidirectional Transformer*	4, 7, 10, 13, 15, 16, 18, 19, 20, 26, 39, 45, 49, 62, 63, 67, 68, 69, 70, 77, 88, 89, 91, 93
GPT	Generative Pretrained Transformer	15, 69, 70, 71
IA	Inteligência Artificial	30
IEEE	Instituto De Engenheiros Eletricistas E Eletrônicos	27, 29, 37
IR	*Isotonic Regression*	57, 58, 85
ISP	Instituto De Segurança Pública	2
LLM	*Large Language Model*	4, 5, 31, 34, 35, 69, 71, 94
LSTM	*Long Short-Term Memory*	33, 62, 69, 70, 71, 88, 89
MCE	*Maximum Calibration Error*	76
MLM	Masked Language Modeling	14
NLL	*Negative Log-Likelihood*	23, 24, 25, 57, 76
NSP	Next Sentence Prediction	14
ODS	Objetivos De Desenvolvimento Sustentável	3
ONGS	Organizações Não Governamentais	32
ONU	Organização Das Nações Unidas	3
PLMS	Modelos De Linguagem Pré-Treinados	13, 14, 15

PLN	Processamento De Linguagem Natural	7, 13, 14, 15, 18, 19, 20, 25, 27, 30, 32, 34, 37, 47
PRISMA	*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*	29, 31
REN	Reconhecimento De Entidades Nomeadas	3, 7, 8, 9, 10, 11, 15, 18, 19, 20, 21, 22, 23, 25, 26, 27, 30, 31, 32, 33, 37, 38, 40, 45, 46, 54, 71, 89, 94
RNNS	Redes Neurais Recorrentes	8, 13, 14
SVM	*Support Vector Machines*	3
T5	Text-To-Text Transfer Transformer	15
TF-IDF	*Term Frequency-Inverse Document Frequency*	33
TS	*Temperature Scaling*	57, 58

Declaração de Uso de Ferramentas de Inteligência Artificial

O autor declara que utilizou ferramentas de Inteligência Artificial generativa exclusivamente como apoio à produção textual desta dissertação, incluindo sugestões de redação, revisão linguística, esclarecimento de conceitos técnicos e organização de ideias. Todas as decisões conceituais, análises, interpretações, escolhas metodológicas, resultados experimentais e conclusões apresentadas são de responsabilidade integral do autor. Nenhum conteúdo gerado por ferramentas de IA foi incorporado sem revisão crítica. O uso dessas ferramentas ocorreu em conformidade com princípios de integridade acadêmica, transparência e responsabilidade, e não substituiu o trabalho intelectual necessário à elaboração desta pesquisa.

Capítulo I Introdução

I.1 Contextualização

No contexto brasileiro, a violência é um problema persistente, afetando de maneira significativa a população, especialmente no Estado do Rio de Janeiro. Conforme os dados apresentados pelo Instituto de Pesquisa Econômica Aplicada [2024], a taxa de homicídios quase triplicou ao longo dos últimos 30 anos. Embora os cinco anos mais recentes tenham mostrado uma tendência de queda, essa redução não compensa a acentuada elevação registrada nas décadas anteriores.

A mesma pesquisa revela que, entre 1995 e 2024, ocorreram aproximadamente 1,3 milhão de homicídios intencionais no Brasil. Apenas em 2024, foram registrados cerca de 38.772 homicídios, representando uma diminuição de 5% em comparação a 2023. Ainda assim, os números mantêm o país entre aqueles com os mais elevados índices de violência do mundo.

No cenário latino-americano, o problema torna-se ainda mais evidente. Dados do Banco Interamericano de Desenvolvimento (BID), conforme apresentados por Perez et al. [2024], mostram que a América latina, embora concentre apenas 9% da população mundial, responde por aproximadamente 33% de todos os homicídios registrados no planeta. Nos últimos cinco anos, a média anual foi de 117 mil homicídios.

Do total de 144 mil homicídios registrados anualmente na América latina, o Brasil é responsável, em média, por cerca de 65 mil ocorrências, o que corresponde a 45% dos casos da região. Esse dado torna-se ainda mais expressivo quando considerado em perspectiva global: embora o país concentre apenas cerca de 3,6% da população mundial, responde por aproximadamente 18% de todos os homicídios registrados no planeta. Esses números evidenciam não apenas a gravidade do fenômeno da violência no contexto latino-americano, mas também o papel central desempenhado pelo Brasil na composição desse cenário, consolidando a região como a mais violenta do mundo em termos absolutos de letalidade.

No âmbito estadual, o Rio de Janeiro apresenta um quadro particularmente sensível. A recorrente escalada dos índices de violência no estado revela fragilidades estruturais da Segurança Pública, relacionadas tanto à limitação de recursos materiais quanto à necessidade de investimentos contínuos na formação e capacitação das forças policiais. A superação desse cenário demanda um esforço governamental articulado, capaz de integrar políticas públicas de prevenção, repressão

qualificada e inteligência, além da adoção de estratégias baseadas em evidências empíricas e análise sistemática de dados.

Esse diagnóstico é corroborado pelo estudo apresentado por Centro de Estudos de Segurança e Cidadania [2025], elaborado pelo Instituto de Segurança Pública (ISP), que aponta, no ano de 2025, um aumento em todos os principais indicadores de violência no Estado do Rio de Janeiro. Tal crescimento contribui de forma significativa para o agravamento da sensação de insegurança da população e reforça a centralidade da atuação do crime organizado como fator estruturante na dinâmica da violência contemporânea no país. Nesse contexto, torna-se evidente a necessidade de aprimorar mecanismos de monitoramento, organização e análise das informações criminais, de modo a subsidiar ações mais eficazes de enfrentamento à criminalidade.

Apesar disso, o investimento em inteligência e tecnologia pode otimizar recursos, reduzir custos e melhorar a eficiência das ações de segurança. Uma das frentes mais promissoras é a organização avançada dos dados, permitindo extrair informações estratégicas capazes de apoiar a formulação e a execução de políticas de combate à violência. O uso adequado desses dados possibilita uma compreensão mais profunda do fenômeno criminal e contribui para o desenvolvimento de medidas mais eficazes de prevenção e repressão.

O Disque Denúncia é uma central de atendimento especializada em auxiliar a população que vivencia ou presencia ações criminosas. Os relatos recebidos por meio de ligações anônimas são repassados às autoridades competentes com rapidez e eficácia, e as denúncias também podem ser registradas diretamente por meio de seu aplicativo. Apesar de sua relevância, não se trata de um órgão governamental, mas sim de uma iniciativa do Instituto MovRio, uma entidade privada e sem fins lucrativos, financiada integralmente por empresas por meio de parcerias, contribuições e convênios [Disque Denúncia, 2023]. esse serviço desempenha um papel central no ecossistema de segurança pública ao atuar como um canal confiável e acessível para o recebimento de informações sobre práticas criminosas. Desde sua criação, em 1995, o serviço consolidou-se como uma das principais fontes de dados sobre a dinâmica do crime no Estado do Rio de Janeiro, reunindo milhões de notificações enviadas anonimamente pela população. Esses registros, estruturados a partir de descrições livres fornecidas pelos denunciantes, constituem o que este trabalho denomina *relatos criminais*: narrativas espontâneas que descrevem eventos, locais, indivíduos e organizações envolvidos em situações suspeitas ou ilícitas.

Nesse cenário, os avanços tecnológicos, aliados à crescente disponibilidade de dados na área de segurança pública, tornam possível o desenvolvimento de modelos e métodos capazes de extrair informações relevantes para auxiliar na elucidação de práticas criminosas. No caso específico dos relatos do Disque Denúncia, além do texto livre (frequentemente informal, ruidoso e de difícil processamento), esses registros são acompanhados por metadados estruturados, como tipo de ocorrência,

localização e data. Esses atributos fornecem um contexto adicional valioso, que pode ser explorado para enriquecer a interpretação dos relatos e reduzir ambiguidades inerentes à linguagem natural. Assim, a integração entre dados textuais não estruturados e metadados estruturados emerge como uma abordagem promissora para ampliar a capacidade analítica dos sistemas de inteligência em segurança pública. Nesse contexto, este trabalho propõe um método para analisar relatos criminais do Disque Denúncia e identificar automaticamente locais, pessoas e organizações mencionadas nos textos por meio de técnicas de Reconhecimento de Entidades Nomeadas (REN), explorando de forma conjunta o conteúdo textual e as informações contextuais disponíveis.

Além dos desafios linguísticos e contextuais, o cenário investigado é caracterizado pela coexistência de dois regimes distintos de dados: um conjunto reduzido de relatos manualmente anotados e um volume significativamente maior de dados não rotulados. Essa assimetria impõe limitações ao uso de abordagens supervisionadas tradicionais e motiva o uso de estratégias semissupervisionadas capazes de explorar ambos os conjuntos de forma integrada, ampliando a cobertura dos dados e o potencial de generalização dos modelos.

I.2 Motivação

Relatos criminais provenientes do Disque Denúncia constituem uma fonte rica de informação para apoio à segurança pública, mas apresentam desafios significativos para métodos de processamento de linguagem natural, em especial para o reconhecimento de entidades nomeadas. Esses textos são marcados por linguagem informal, variações lexicais, ambiguidades e inconsistências estruturais, o que dificulta a aplicação direta de modelos desenvolvidos em domínios mais formais.

Trabalhos anteriores têm explorado o uso de técnicas de aprendizado de máquina nesse contexto. da Silva and Melo [2020], por exemplo, utilizam modelos tradicionais, como *Random Forest*, *Support Vector Machines* (SVM) e *Naive Bayes*, para classificar textos associados ao Disque Denúncia. De forma complementar, Figueiredo et al. [2023] investigam o reconhecimento de entidades nomeadas a partir de um *corpus* anotado, empregando *word embeddings* específicos ao domínio.

Entretanto, tais abordagens baseiam-se exclusivamente em conjuntos de dados rotulados de menor escala, o que limita sua capacidade de generalização e dificulta a aplicação em cenários reais, nos quais predominam grandes volumes de dados não anotados. Além disso, essas abordagens não exploram de forma sistemática os metadados estruturados associados aos relatos, nem o potencial de modelos de linguagem de grande porte ajustados ao domínio.

Adicionalmente, a proposta alinha-se ao Objetivos de Desenvolvimento Sustentável (ODS) 16, estabelecido pela Organização das Nações Unidas (ONU), ao contribuir para o fortalecimento de práticas de segurança pública orientadas por dados. Ao possibilitar a extração estruturada de informações presentes em relatos criminais informais, por meio do reconhecimento de entidades

nomeadas, o trabalho favorece a organização e a interpretação automatizada desses dados, podendo atuar como etapa intermediária em sistemas de apoio à análise e à tomada de decisão no contexto da segurança pública.

Nesse contexto, observa-se uma lacuna na literatura quanto ao uso integrado de modelos de linguagem de grande porte, aprendizado semissupervisionado e metadados estruturados para o reconhecimento de entidades em linguagem informal, especialmente em aplicações voltadas à segurança pública.

I.3 Objetivos

Este trabalho tem como objetivo principal desenvolver uma metodologia para o reconhecimento de entidades nomeadas em relatos criminais informais do Disque Denúncia, considerando um cenário caracterizado pela coexistência de um conjunto reduzido de dados rotulados e um volume substancialmente maior de dados não anotados. A proposta busca explorar de forma integrada esses dois regimes de dados, combinando modelos de linguagem de grande porte com estratégias semissupervisionadas e o uso de metadados estruturados como fonte adicional de contexto.

Especificamente, pretende-se: (i) identificar automaticamente entidades do tipo Pessoa, Local e Organização em textos informais; (ii) aplicar técnicas de ajuste fino em modelos baseados em *Large Language Model* (LLM) a partir de um *corpus* anotado; (iii) investigar o uso de pseudo-rotulagem para expandir o conjunto de treinamento a partir de dados não rotulados; e (iv) avaliar o impacto da incorporação de metadados no desempenho dos modelos, com ênfase no modelo *Generalist and Lightweight Model for Named Entity Recognition using Bidirectional Transformer* (GLiNER).

Adicionalmente, o trabalho busca contribuir para a estruturação e análise de dados provenientes de relatos criminais, apoiando o desenvolvimento de ferramentas de inteligência voltadas à segurança pública, ao mesmo tempo em que preserva o anonimato dos denunciantes.

Dessa forma, o problema central desta pesquisa pode ser formulado como:

> *Como reconhecer automaticamente entidades nomeadas (pessoas, locais e organizações) em relatos criminais escritos em linguagem informal, explorando de forma integrada o conteúdo textual, os metadados estruturados e a complementaridade entre dados rotulados e não rotulados, de modo a aprimorar o desempenho de modelos de linguagem ajustados ao domínio da segurança pública?*

I.4 Visão Geral da Abordagem

A metodologia proposta neste trabalho baseia-se na integração de modelos de linguagem de grande porte com estratégias semissupervisionadas, considerando um cenário caracterizado pela

coexistência de dados rotulados em pequena escala e um volume significativamente maior de dados não anotados provenientes do Disque Denúncia.

Inicialmente, os dados são organizados em dois conjuntos complementares: um *corpus* anotado, utilizado para o ajuste fino supervisionado dos modelos, e um conjunto ampliado de relatos não rotulados, explorado por meio de técnicas de pseudo-rotulagem. Esse processo permite expandir progressivamente o conjunto de treinamento, mitigando a escassez de dados anotados e ampliando a capacidade de generalização dos modelos.

Os modelos de linguagem são então ajustados ao domínio dos relatos criminais, incorporando características específicas da linguagem informal presente nesses dados. Adicionalmente, os metadados estruturados associados às denúncias são integrados ao processo de modelagem como fonte de contexto, contribuindo para a redução de ambiguidades e para o enriquecimento das representações aprendidas.

Por fim, o desempenho dos modelos é avaliado por meio de experimentos comparativos com diferentes configurações e *baselines*, considerando métricas tradicionais de reconhecimento de entidades nomeadas. Essa avaliação busca analisar, de forma sistemática, o impacto da pseudo-rotulagem e da incorporação de metadados na qualidade das predições.

Essa abordagem permite investigar, de forma integrada, os efeitos da combinação entre modelos de linguagem ajustados ao domínio, aprendizado semissupervisionado e informações contextuais estruturadas na tarefa de reconhecimento de entidades em linguagem informal. Além disso, os resultados obtidos podem contribuir para o desenvolvimento de plataformas de apoio à análise de relatos criminais, ampliando a capacidade de exploração dos dados presentes nos *corpora* do Disque Denúncia e auxiliando na identificação de padrões e relações relevantes para a segurança pública.

I.5 Organização

Esta dissertação está estruturada em cinco capítulos, além desta introdução. O Capítulo II apresenta a fundamentação teórica que embasa o estudo, descrevendo os principais conceitos, técnicas e modelos de aprendizado de máquina e de modelos de linguagem de grande porte (LLM) utilizados ao longo do trabalho. No Capítulo III discutem-se os trabalhos relacionados que contribuíram para o desenvolvimento desta pesquisa, evidenciando abordagens, limitações e avanços presentes na literatura. O Capítulo IV detalha a metodologia adotada para a resolução do problema proposto, descrevendo a preparação dos dados, o processo de modelagem e os modelos de classificação desenvolvidos. O Capítulo V apresenta os experimentos computacionais realizados, detalhando sua execução, análise e interpretação para evidenciar os objetivos específicos e os resultados obtidos. O capítulo também descreve o plano experimental proposto, os testes implementados, seus respectivos critérios de avaliação e os procedimentos utilizados para verificar a efetividade dos métodos

empregados. Além disso, discute-se o desempenho alcançado em cada cenário, contribuindo para a compreensão do impacto das técnicas utilizadas. Por fim, o Capítulo VI reúne as conclusões do estudo e apresenta os próximos passos necessários para a conclusão da pesquisa.

Capítulo II Referencial Teórico

Este capítulo apresenta os fundamentos teóricos necessários para a compreensão dos capítulos seguintes. A Seção II.1 aborda o REN, com seu histórico, esquemas de anotação e suas métricas de avaliação. Em seguida, a Seção II.2 trata dos conceitos vinculados aos modelos *Transformers* e descreve os modelos de linguagem pré-treinados, com destaque para o GLiNER, selecionado para aplicação nesta dissertação e a técnica de ajuste fino. Por fim, a Seção II.3 apresenta a técnica de pseudo-rotulação empregada para ampliar a capacidade de reconhecimento das entidades e descreve os métodos de calibração aplicados aos modelos gerados, bem como as métricas utilizadas para avaliar sua efetividade.

II.1 Reconhecimento de Entidades Nomeadas (REN)

II.1.1 Definição e Histórico

A classificação de *tokens* é uma tarefa fundamental do Processamento de Linguagem Natural (PLN), cujo objetivo é atribuir a cada palavra de um texto uma categoria pré-definida. Essa atividade é essencial em aplicações como REN, rotulagem morfossintática (*POS tagging*) e análise de sentimentos. Para executá-la de forma eficaz, os modelos precisam compreender tanto o contexto local quanto o global das sentenças, o que representa um desafio especial em línguas morfologicamente complexas, como o português.

Com os avanços em *Deep Learning*, especialmente com o surgimento dos modelos baseados na arquitetura *Transformer*, como o *Bidirectional Encoder Representations from Transformers* (BERT) [Devlin et al., 2019], a classificação de *tokens* atingiu novos níveis de desempenho. A capacidade do *Transformer* de processar sequências inteiras em paralelo e capturar dependências de longo alcance permitiu melhorias significativas em tarefas dependentes de contexto, incluindo o REN.

Técnicas de transferência de aprendizado e ajuste fino também ganharam destaque, pois possibilitam adaptar modelos pré-treinados em grandes *corpora* para domínios específicos, mesmo quando se dispõe de conjuntos limitados de dados anotados [Souza et al., 2020].

O REN visa identificar e classificar, dentro de um texto, entidades como pessoas, organizações, locais, medicamentos, horários, proteínas e procedimentos clínicos. Trata-se de uma etapa crítica em diversos *pipelines* do PLN, como a resposta a perguntas, a extração de relações, a recuperação de

informações e a modelagem de tópicos. Embora o estudo do REN exista há décadas, sua aplicação com modelos de redes neurais profundas tornou-se mais expressiva apenas nos últimos anos.

A primeira tarefa de REN foi introduzida por Grishman and Sundheim [1996] na *Sixth Message Understanding Conference*. As abordagens iniciais utilizavam regras lexicais, características ortográficas e ontologias. Posteriormente, modelos estatísticos como os Modelos Ocultos de Markov [Baum et al., 1970] e os Campos Aleatórios Condicionais [Lafferty et al., 2001] passaram a dominar a área.

O trabalho de Collobert et al. [2011] marcou a transição para redes neurais com baixa dependência de engenharia de atributos, abrindo caminho para representações distribuídas e arquiteturas mais robustas. Desde então, diversos modelos foram propostos, adotando Redes Neurais Recorrentes (RNNs) e mecanismos de incorporação de caracteres, subpalavras ou *embeddings* pré-treinados.

Modelos baseados em *Transformers* consolidaram-se como padrão de estado da arte no REN [Labusch et al., 2019; Jeong and Kim, 2022; Vacareanu et al., 2024; Shi and Kimura, 2024]. O número de publicações relacionadas ao tema cresceu de forma exponencial desde 2018, refletindo a ampla adoção dessas arquiteturas.

II.1.2 Esquemas de anotação

O processo de reconhecimento de entidades nomeadas requer a representação explícita das fronteiras das entidades no texto. Essa representação é tradicionalmente realizada por meio de esquemas de anotação que convertem sequências de palavras em sequências de rótulos estruturados, indicando o início, o interior e o fim de cada entidade. Dentre os esquemas mais utilizados na literatura, destacam-se o *Begin-Inside-Outside* (BIO), o *Begin, Inside, Last, Outside, Unit* (BILOU) e, mais recentemente, o SpanLabel; este último está intimamente associado a abordagens baseadas em spans em modelos de aprendizado profundo [Ratinov and Roth, 2009; Yadav and Bethard, 2019; Li et al., 2020].

Esquema BIO (*Begin, Inside, Outside*)

O esquema BIO é um dos métodos mais tradicionais de anotação em tarefas de REN. Nesse formato, cada *token* recebe uma etiqueta que indica se ele:

- **B-XXX**: início de uma entidade;
- **I-XXX**: continuação da mesma entidade;
- **O**: *token* fora de qualquer entidade.

Por exemplo, na sentença "Cristiano Ronaldo nasceu em Aveiro", a sequência de rótulos no formato BIO seria:

Cristiano Ronaldo nasceu em Aveiro

`[B-PESSOA Cristiano] [I-PESSOA Ronaldo] [O nasceu] [O em] [B-LOCAL Aveiro]`

Esse esquema, inicialmente popularizado por tarefas de *chunking* e posteriormente adaptado para REN, oferece uma codificação simples e eficiente para modelos baseados em sequências [Ratinov and Roth, 2009]. Não obstante, apresenta limitações na representação de fronteiras explícitas, uma vez que não diferencia o fim de uma entidade, apenas o início e sua continuação.

Esquema BILOU (*Begin, Inside, Last, Outside, Unit*)

O esquema BILOU é uma extensão do BIO que adiciona maior granularidade semântica, permitindo uma distinção explícita entre o início, o meio e o fim de uma entidade. Ele introduz dois novos rótulos:

- **L-XXX**: último *token* da entidade;
- **U-XXX**: entidade composta por um único *token*.

Assim, a mesma sentença anterior seria rotulada como:

`[B-PESSOA Cristiano] [L-PESSOA Ronaldo] [O nasceu] [O em] [U-LOCAL Aveiro]`

O esquema BILOU oferece vantagens teóricas em relação ao BIO, pois explicita o limite final das entidades, o que facilita o aprendizado de fronteiras em modelos baseados em redes neurais recorrentes e *Transformers* [Lample et al., 2016]. Estudos empíricos demonstram que esse detalhamento adicional contribui para uma melhor precisão na detecção e segmentação de entidades, reduzindo ambiguidades em textos com múltiplas entidades adjacentes [Yadav and Bethard, 2019].

Esquema SpanLabel

Com o avanço de arquiteturas baseadas em *Transformers* e abordagens de extração por spans, emergiu o esquema SpanLabel, que se distingue conceitualmente dos modelos tokenizados (BIO e BILOU). Nesse caso, o texto não é rotulado *token* a *token*; em vez disso, são definidos diretamente os intervalos de início e fim (*spans*) das entidades no texto. Cada entidade é representada como um triplo no formato:

(Índice do vetor do Inicio da Entidade,Índice do vetor do Fim da Entidade,Tipo da Entidade)

Por exemplo:

Cristiano Ronaldo nasceu em Aveiro
`[(0,17, PESSOA), (28,34, LOCAL)]`

Essa abordagem é particularmente útil em contextos de entidades sobrepostas, em que diferentes entidades podem se interceptar parcial ou completamente no texto [Li et al., 2020]. Além disso, o esquema SpanLabel se integra naturalmente a modelos neurais modernos, como o GLiNER, que opera sobre pares de índices em vez de rótulos sequenciais, favorecendo a representação contextual e a extração de múltiplas entidades simultaneamente.

II.1.3 Métricas de avaliação

A avaliação de modelos de aprendizado supervisionado, em especial aqueles aplicados ao REN, depende de métricas capazes de quantificar o desempenho do sistema. Entre as medidas mais utilizadas estão a Precisão, o *Revocação* e o *F1*, que, em conjunto, permitem uma análise equilibrada da qualidade das predições realizadas pelo modelo [Sokolova and Lapalme, 2009; Powers, 2020].

A Precisão expressa a proporção de resultados corretos entre todas as predições positivas feitas pelo modelo. Em outras palavras, indica quão confiável é quando decide que um elemento pertence a uma determinada classe. Um modelo com alta precisão comete poucos erros de classificação incorreta, o que é desejável em aplicações sensíveis a falsos positivos, como a detecção de fraudes ou diagnósticos clínicos automáticos [Powers, 2020].

A Precisão quantifica a proporção de entidades previstas corretamente entre todas as entidades que o modelo identificou. Sua fórmula é dada pela Equação II.1:

$$\text{Precisão} = \frac{\text{Entidades corretamente previstas}}{\text{Entidades previstas pelo modelo}} = \frac{TP}{TP + FP} \tag{II.1}$$

Na equação acima:

- TP – Verdadeiros Positivos (entidades corretamente reconhecidas e classificadas);
- FP – Falsos Positivos (entidades incorretamente reconhecidas ou classificadas).

Por outro lado, o *Revocação* (ou Sensibilidade) mede a capacidade do modelo de identificar corretamente todos os elementos que realmente pertencem a uma classe. Essa métrica é especialmente importante em contextos nos quais é mais prejudicial deixar de identificar uma ocorrência verdadeira do que realizar uma previsão incorreta, como no caso da identificação de entidades em textos criminais ou médicos, em que a omissão de uma entidade relevante pode comprometer análises posteriores [Goutte and Gaussier, 2005].

Quando temos bases de dados desbalanceadas, há uma divergência entre a precisão e a *revocação*, pois aumentar um tende a reduzir o outro. Por exemplo, um modelo com alta precisão pode

classificar entidades mais fáceis, o que reduz o *revocação*. E um modelo que tenta identificar todas as possíveis instâncias verdadeiras e pode aumentar o *revocação*, mas à custa de uma precisão menor.

O *Revocação* mede a proporção de entidades reais corretamente identificadas pelo modelo:

$$\text{Revocação} = \frac{\text{Entidades corretamente previstas}}{\text{Entidades reais no conjunto de dados}} = \frac{TP}{TP + FN} \tag{II.2}$$

Onde:

- TP – Verdadeiros Positivos;
- FN – Falsos Negativos (entidades presentes no texto que o modelo não reconheceu).

No cenário acima, o *F1* se apresenta como uma métrica de síntese amplamente aceita na literatura de aprendizado de máquina e processamento de linguagem natural [Yadav and Bethard, 2019]. Essa métrica combina de forma harmônica o *revocação* e a precisão para oferecer uma visão melhor do desempenho do modelo. Assim, o *F1* penaliza fortemente casos em que uma das duas métricas é muito baixa, garantindo que o modelo mantenha um desempenho equilibrado entre identificar corretamente e não cometer excessos de classificação [Li et al., 2020].

O *F1* corresponde à média harmônica entre Precisão e Revocação:

$$\text{F1} = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} = \frac{2TP}{2TP + FP + FN} \tag{II.3}$$

Com:

- TP – Verdadeiros Positivos;
- FP – Falsos Positivos;
- FN – Falsos Negativos.

Formas de avaliação: *token-level* e *entity-level*

No contexto do REN, o desempenho do modelo pode ser avaliado de duas formas principais: ao nível de *token* (*token-level*) e ao nível de entidade (*entity-level*). Embora ambas utilizem métricas como Precisão, *Revocação* e *F1*, elas diferem na forma como o acerto é contabilizado.

Na avaliação ao nível de *token-level*, cada palavra (ou subpalavra, no caso de modelos baseados em *Transformers*) é avaliada individualmente. Essa abordagem considera o acerto de rótulos *token* por *token*, sendo mais tolerante a erros parciais. Por exemplo, se o modelo rotular corretamente

parte de uma entidade, ele ainda receberá crédito parcial. Essa forma de avaliação tende a apresentar valores de *F1* ligeiramente mais altos, uma vez que pequenos erros de segmentação não penalizam fortemente o desempenho.

Já a avaliação ao nível de *entity-level* considera uma entidade como corretamente identificada apenas se todos os *tokens* que a compõem forem previstos corretamente com o mesmo rótulo. Caso contrário, a predição é considerada incorreta. Essa abordagem reflete de forma mais fiel à utilidade prática do modelo, especialmente em contextos nos quais é necessário reconhecer entidades completas, como nomes de pessoas, locais ou organizações. Como destacado no exemplo abaixo:

Tabela II.1: Comparação entre abordagens de avaliação em REN

Critério	Avaliação em *Token-Level*	Avaliação em *Entity-Level*
Unidade de análise	Cada *token* ou subpalavra é avaliado individualmente.	A entidade completa é avaliada como uma única unidade.
Tolerância a erros	Permite acertos parciais em entidades parcialmente reconhecidas.	Exige correspondência exata de início, fim e tipo da entidade.
Complexidade da avaliação	Métricas simples e interpretação direta.	Avaliação mais rigorosa e estruturalmente complexa.
Tendência dos resultados	Produz valores de *F1* geralmente mais elevados.	Produz valores de *F1* mais conservadores e realistas.
Uso recomendado	Análises exploratórias e diagnósticos intermediários.	Avaliações finais e comparação formal entre modelos.

"Cristiano Ronaldo nasceu em Aveiro."

A anotação de referência (dado rotulado) é:

```
[(0,17,PESSOA), (28,34,LOCAL)]
```

onde "Cristiano Ronaldo" é uma entidade do tipo Pessoa e "Aveiro" é uma entidade do tipo Local.

Suponha que o modelo tenha previsto apenas "Cristiano" como a entidade Pessoa. No nível de *token-level*, o modelo receberá crédito parcial, pois reconheceu parte da entidade corretamente. Entretanto, no nível de *entity-level*, essa predição será considerada incorreta, uma vez que a entidade completa "Cristiano Ronaldo" não foi identificada integralmente.

Esse exemplo demonstra que a avaliação em *entity-level* é mais rigorosa, refletindo com maior precisão o desempenho real do modelo em tarefas de reconhecimento completo de entidades.

II.2 Transformers e Modelos Pré-treinados

As subseções a seguir apresentam os fundamentos teóricos que sustentam o uso de modelos modernos de PLN baseados na arquitetura *Transformer*. Primeiramente, descreve-se o funcionamento dos *Transformers* (Seção II.2.1) e os avanços que essa arquitetura introduziu em relação aos modelos sequenciais tradicionais. Em seguida, discutem-se os Modelos de Linguagem Pré-Treinados (PLMs) (Seção II.2.2) e o que essa arquitetura introduziu em relação aos modelos sequenciais tradicionais que se consolidaram como estado da arte em tarefas de PLN ao combinar pré-treinamento em larga escala com técnicas de ajuste fino. Por fim, a Seção II.2.3 apresenta o modelo GLiNER, adotado neste trabalho. Essa Seção estabelece a base conceitual necessária para compreender os métodos empregados neste trabalho, enquanto a técnica de ajuste fino aplicada ao GLiNER é detalhada posteriormente na Seção II.2.4.

II.2.1 Transformers

O modelo *Transformer*, introduzido por Vaswani et al. [2017], foi concebido como uma alternativa aos modelos baseados em RNNs e em redes convolucionais, que até então eram predominantes em tarefas de tradução automática e modelagem de sequência. Sua principal inovação é o uso exclusivo de mecanismos de atenção, eliminando a necessidade de recorrência. Isso permitiu maior paralelização e eficiência no treinamento, sem perda de desempenho.

A arquitetura do *Transformer* (Figura II.1) é composta por dois blocos simétricos:

- **Encoder**: responsável por codificar a sequência de entrada e gerar representações contextuais ricas;
- **Decoder**: responsável por gerar a sequência de saída, utilizando mecanismos de autoatenção e atenção sobre o encoder.

Cada bloco é formado pela combinação de:

- mecanismos de atenção multi-cabeça (*multi-head attention*);
- redes *feedforward*;
- conexões residuais e normalização.

Como o modelo não utiliza recorrência, incorpora-se o conceito de *positional encodings*, vetores senoidais que introduzem informações de ordem na sequência, permitindo ao modelo interpretar relações posicionais entre *tokens*.

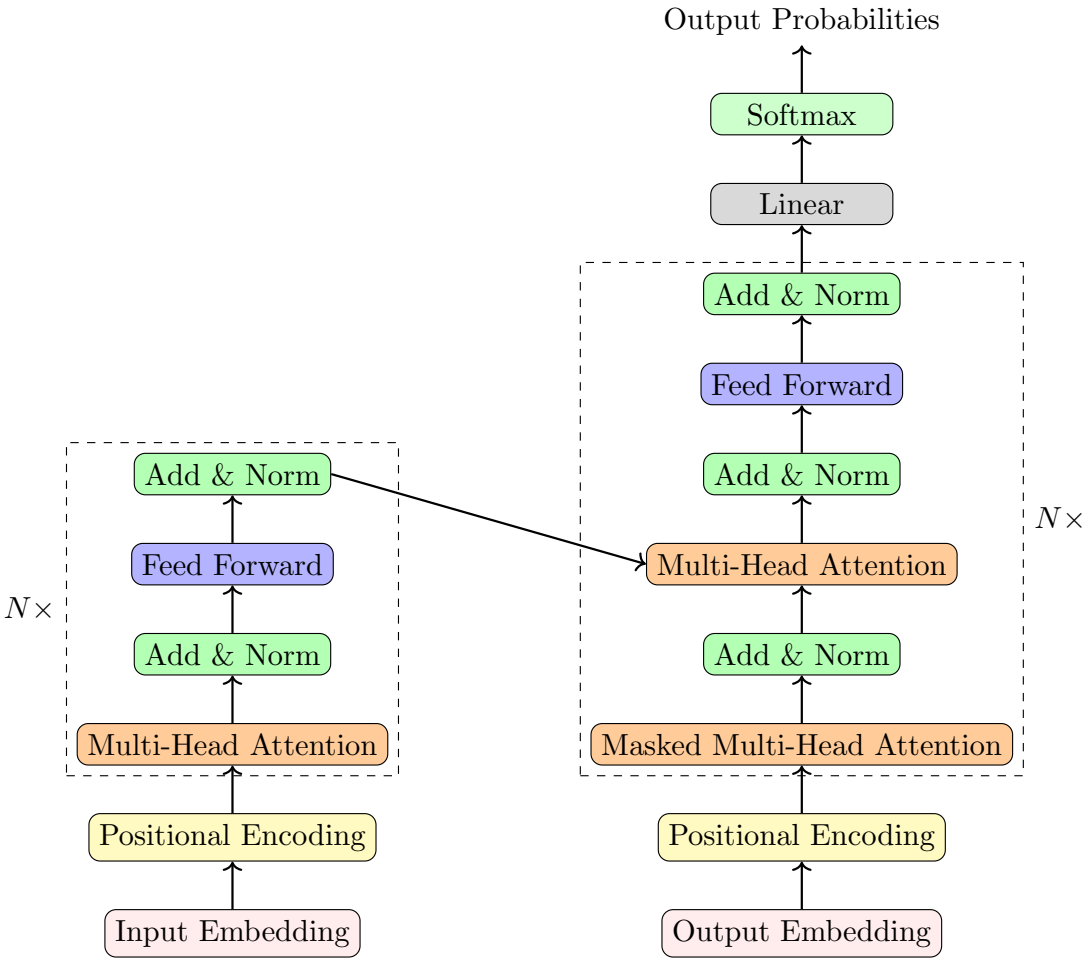


Figura II.1: Arquitetura do modelo *Transformer*, baseado na proposta Vaswani et al. [2017]. O modelo é composto por um codificador (*encoder*) e um decodificador (*decoder*), ambos formados por múltiplas camadas que incluem mecanismos de atenção multi-cabeças (*multi-head attention*), normalização residual (*add & norm*) e redes *feedforward*. Essa estrutura permite o aprendizado de representações contextuais eficientes em tarefas de processamento de linguagem natural.

Essa estrutura altamente paralelizável tornou possível o treinamento de modelos muito maiores, capacitando o *Transformer* a capturar dependências de longo alcance, superar limitações das RNNs e alcançar melhores resultados em tarefas como tradução, inferência textual e classificação [Vaswani et al., 2017].

Com o surgimento do BERT, o *Transformer* passou a ser utilizado como base para modelos de linguagem pré-treinados, que introduziram objetivos de aprendizado como o Masked Language Modeling (MLM) e o Next Sentence Prediction (NSP), revolucionando o paradigma de ajuste fino no PLN [Devlin et al., 2019].

II.2.2 Modelos de Linguagens Pré-Treinados

Os PLMs são redes neurais de larga escala baseadas na arquitetura *Transformer*, treinadas inicialmente sobre grandes volumes de texto utilizando tarefas auto-supervisionadas, como a previsão de palavras mascaradas (MLM) ou a predição de próximas palavras [Touvron et al., 2023]. Es-

ses modelos aprendem padrões linguísticos complexos e generalizáveis, podendo ser posteriormente adaptados a diversas tarefas específicas por meio do ajuste fino.

Os modelos baseados na arquitetura *Transformer* podem ser classificados em três categorias principais: *encoder-only, decoder-only e encoder-decoder*.

- **Encoder-only**: como o BERT, ideal para tarefas discriminativas (classificação, REN);
- **Decoder-only**: como o Generative Pretrained Transformer (GPT), orientados para tarefas generativas;
- **Encoder-decoder**: como o Text-To-Text Transfer Transformer (T5), adequados para tarefas de transformação de sequência, como tradução.

Modelos *encoder-only* como o BERT são bidirecionais e capturam o contexto à esquerda e à direita de cada *token*, sendo particularmente eficazes em tarefas como classificação de sentenças e REN [Devlin et al., 2019]. Em contraste, modelos *decoder-only* ou *encoder-decoder* funcionam de forma causal (unidirecional ou sequencial), tornando-os ideais para tarefas gerativas [Achiam et al., 2023].

O sucesso dos PLMs deve-se, em grande parte, à combinação entre:

- pré-treinamento massivo em grandes *corpora*;
- poder expressivo das arquiteturas *Transformer*;
- capacidade de adaptação por ajuste fino.

Como consequência, os PLMs tornaram-se o núcleo das aplicações modernas em PLN, estabelecendo novos padrões de desempenho em tarefas como tradução automática, reconhecimento de entidades nomeadas, sumarização e análise de sentimentos.

II.2.3 GliNER

O GLiNER, proposto por Zaratiana [2023], foi desenvolvido como uma solução moderna e eficiente para o REN em cenários de vocabulário aberto (*open-vocabulary*). Diferentemente dos modelos tradicionais baseados em esquemas como BIO ou BILOU, o GLiNER utiliza descrições textuais das classes de entidades, permitindo que novas categorias sejam reconhecidas sem a necessidade de re-treinar o modelo. Essa característica o torna particularmente útil em domínios com rótulos altamente dinâmicos ou com baixa disponibilidade de dados anotados.

O modelo utiliza um *encoder* bidirecional baseado na arquitetura *Transformer* (como BERT ou *Decoding-enhanced BERT with Disentangled Attention* (DeBERTa)) para gerar representações

contextuais dos *tokens*. Em paralelo, as descrições textuais dos tipos de entidades também são codificadas no mesmo espaço vetorial. Dessa forma, o reconhecimento é realizado por meio da comparação entre *spans* do texto e as representações dos tipos de entidades, utilizando medidas de similaridade para determinar as correspondências mais prováveis [Zilliz, 2024].

A arquitetura do GLiNER combina um modelo de linguagem bidirecional, como BERT, com um módulo de codificação de rótulos que processa as descrições das entidades [Nguyen, 2024]. O alinhamento entre *spans*(segmentos contínuos) e rótulos ocorre por meio de similaridade vetorial, o que possibilita identificar classes não vistas durante o treinamento ou com poucos exemplos [Keraghel et al., 2024]. O treinamento do GLiNER garante maior cobertura e capacidade de generalização. Essa abordagem também é explorada em variações multitarefa, que ampliam o uso do modelo para tarefas de extração de relações [Stepanov et al., 2024].

A arquitetura do modelo GLiNER, descrita na Figura II.2, destaca as representações das entidades que passam por uma rede *feedforward* (*FeedForward Network*), que refina os vetores de cada tipo de entidade. As representações das palavras do texto são enviadas para uma camada de representação de spans, que calcula vetores para cada possível span.

Por fim, o modelo calcula uma pontuação de correspondência (*matching score*) entre as representações das entidades e as representações dos spans, utilizando o produto escalar seguido de uma função sigmoide para gerar uma probabilidade.

Etapas de funcionamento do GLiNER.

De acordo com Zaratiana [2023], o processo de reconhecimento de entidades no GLiNER ocorre em quatro etapas principais:

1. **Codificação do texto de entrada:** A sentença é processada por um modelo de linguagem bidirecional baseado em *Transformers*, que gera *embeddings* contextuais para cada *token*, incorporando informações semânticas e posicionais.

2. **Codificação dos tipos de entidade:** Cada tipo de entidade (por exemplo, Pessoa, Organização e Local) é transformado em uma representação vetorial a partir de sua descrição textual, produzindo um conjunto de *embeddings* semanticamente interpretáveis.

3. **Geração das representações de spans:** As representações dos *tokens* são combinadas para formar possíveis *spans* do texto. Esses spans são refinados por uma camada *feedforward*, que modela interações locais entre os *tokens*.

4. **Cálculo de similaridade e classificação:** Para cada par *(tipo de entidade, span)*, o modelo calcula uma pontuação de similaridade usando o produto escalar entre os *embeddings*

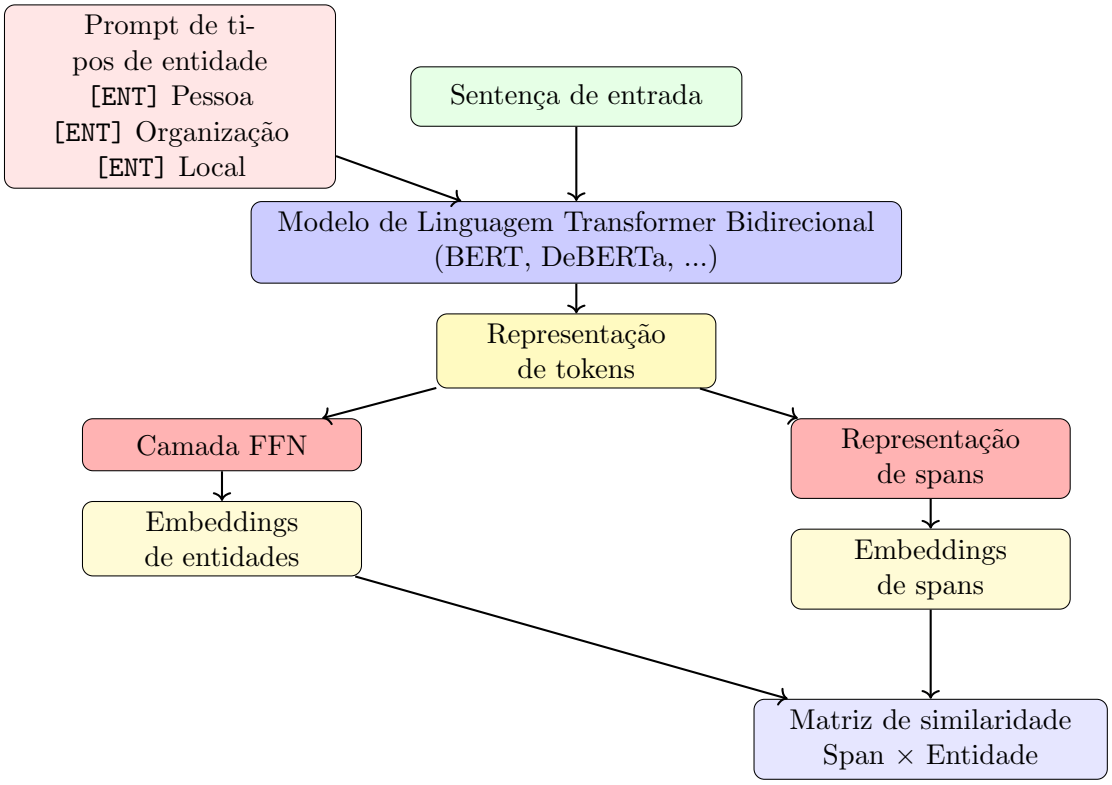


Figura II.2: Arquitetura do modelo *GliNER*, adaptado de Zaratiana [2023]. O modelo baseia-se em uma arquitetura *Transformer* bidirecional (como BERT ou DeBERTa) e utiliza um mecanismo de *prompting* de tipos de entidades, representados como *embeddings* específicos. A sentença de entrada é processada pelo modelo, que gera representações vetoriais posteriormente combinadas em camadas de *feedforward* e na camada de representação de *spans*. Em seguida, é construída uma matriz de similaridade entre os tipos de entidades e os *spans*, a partir da qual o modelo identifica as entidades nomeadas de forma contextualizada.

correspondentes. Essa pontuação é normalizada por uma função sigmoide, produzindo uma probabilidade de associação entre o *span* a determinado tipo de entidade.

Para ilustrar o funcionamento, considere a frase:

"Cristiano Ronaldo nasceu em Aveiro."

Durante o processamento:

1. O texto é tokenizado e cada palavra recebe um vetor contextual:

$$[\text{Cristiano, Ronaldo, nasceu, em, Aveiro}] \rightarrow [h_1, h_2, h_3, h_4, h_5]$$

2. As descrições de entidades (Pessoa e Local) são codificadas em *embeddings* específicos.

3. O modelo gera possíveis *spans*, como "Cristiano Ronaldo" e "Aveiro".

4. Em seguida, calcula-se a similaridade entre cada *span* e os tipos de entidade, resultando em maior probabilidade para:

$$(\text{Cristiano Ronaldo}, \text{Pessoa}) \quad \text{e} \quad (\text{Aveiro}, \text{Local})$$

5. O modelo então retorna as entidades reconhecidas, representadas como:

$$[(0, 17, \text{B-PER}), (28, 34, \text{B-LOCAL})]$$

O GLiNER representa um avanço significativo na área de REN, combinando baixo custo computacional, adaptabilidade e alto desempenho. Como destacam Zaratiana [2023], "o modelo alcança resultados competitivos em cenários *zero-shot* e *few-shot*, mantendo eficiência em hardware limitado". Isso o torna aplicável na análise de notícias, no monitoramento de redes sociais, na extração de informações biomédicas e na detecção de fraudes [Zilliz, 2024].

No *few-shot learning*, o modelo é inicialmente pré-treinado em grandes corpora ou em múltiplas tarefas de PLN, aprendendo representações linguísticas gerais que podem ser posteriormente adaptadas a novos domínios ou conjuntos de classes com um número muito reduzido de exemplos anotados. Esse paradigma permite especializar o modelo sem a necessidade de um esforço extensivo de rotulagem, explorando o conhecimento adquirido durante o pré-treinamento por meio de *encoders* contextuais baseados em arquiteturas baseadas em modelos bidirecionais do tipo BERT ou DeBERTa, que demonstram elevada capacidade de transferência mesmo em cenários de escassez de dados [Stepanov et al., 2024].

Em contraste, no *zero-shot learning*, o modelo deve lidar com classes ou tipos de entidades que não aparecem explicitamente rotulados durante o treinamento supervisionado. Nessa configuração, o reconhecimento depende exclusivamente de descrições semânticas das classes, geralmente expressas em linguagem natural, e de um espaço de representação compartilhado entre instâncias textuais e rótulos. Trabalhos recentes em extração de informação e REN de vocabulário aberto demonstram que, nessa formulação baseada em *matching* semântico, a tarefa pode ser tratada como um problema de alinhamento ou compatibilidade entre representações, dispensando, nesse tipo de modelo, tanto exemplos anotados adicionais quanto mecanismos de geração autorregressiva [Zaratiana, 2023].

O GLiNER está disponível em diferentes configurações, variando conforme a complexidade e a capacidade computacional: GLiNER *Small* (com aproximadamente 50 milhões de parâmetros), GLiNER *Medium* (90 milhões) e GLiNER *Large* (300 milhões de parâmetros).

Por integrar descrições textuais de entidades com representações contextuais baseadas em *Transformers*, o GLiNER representa um avanço significativo no REN, sobretudo em cenários de voca-

bulário aberto e bases desbalanceadas.

II.2.4 Ajuste Fino

O ajuste fino (*fine-tuning*) é uma técnica amplamente utilizada em PLN para especializar modelos de linguagem pré-treinados na tarefa ou domínio específico. Nessa abordagem, parte do conhecimento adquirido durante o pré-treinamento em grandes *corpora* é preservada, enquanto os parâmetros do modelo são ajustados para refletir as características linguísticas e semânticas do novo contexto [Howard and Ruder, 2018]. Em tarefas de REN, o ajuste fino tem mostrado ser particularmente eficaz, sobretudo quando o conjunto de dados anotados é limitado ou apresenta características distintas do domínio original de treinamento [Devlin et al., 2019].

No contexto do GLiNER, o ajuste fino assume um papel central para melhorar o desempenho do modelo em domínios especializados. O GLiNER utiliza representações vetoriais compartilhadas entre *spans* do texto e descrições de entidades, e o processo de ajuste fino permite refinar esse espaço vetorial para maximizar a separabilidade entre entidades relevantes do domínio. Como consequência, o modelo torna-se mais sensível a variações linguísticas específicas e as relações semânticas presentes no *corpus* anotado.

Durante o ajuste fino, o modelo é otimizado com base na maximização das pontuações de similaridade (*matching score*) entre os *spans* corretos e seus respectivos tipos de entidade, calculadas pelo produto escalar seguido da função sigmoide [Zaratiana, 2023]. Essa otimização é conduzida por funções de perda, como a *binary cross-entropy*, aplicadas a todas as combinações possíveis de pares *(span, entidade)*, o que orienta os parâmetros do modelo a se aproximarem de *spans* corretos de seus rótulos e a afastarem combinações incorretas.

O ajuste fino oferece diversas vantagens práticas:

- adapta o modelo às especificidades linguísticas do domínio alvo, como gírias, abreviações ou estruturas sintáticas recorrentes;
- melhora a precisão em entidades pouco frequentes, especialmente em bases desbalanceadas;
- preserva a capacidade do modelo de reconhecer novas classes, característica fundamental do GLiNER;
- reduz a necessidade de grandes volumes de dados anotados manualmente.

Ao contrário de modelos tradicionais de REN, que exigem a reconstrução completa do classificador sempre que novas classes são adicionadas, o GLiNER permite incorporar descrições textuais de novos tipos de entidades sem alterar sua estrutura. O ajuste fino, portanto, fortalece essa propri-

edade ao aperfeiçoar o espaço vetorial compartilhado, permitindo que o modelo se adapte a novos cenários sem perder sua capacidade de generalização.

Assim, o ajuste fino do GLiNER representa uma estratégia eficiente para maximizar o desempenho do modelo em domínios específicos, mantendo sua versatilidade e capacidade de operar em cenários de vocabulário aberto.

II.3 Outras Técnicas Usadas neste Trabalho

II.3.1 Validação Cruzada

A validação cruzada (*k-fold-cross-validation*) é uma técnica amplamente utilizada para estimar a capacidade de generalização de modelos de aprendizado supervisionado, sendo particularmente relevante em tarefas de REN. Em contextos nos quais os conjuntos de dados anotados são reduzidos ou apresentam distribuição heterogênea entre as entidades, a validação cruzada permite obter avaliações mais estáveis e representativas do desempenho do modelo [Gonzalez Hernandez et al., 2024].

No procedimento tradicional de *k-fold*, o conjunto de dados é particionado em k subconjuntos (*folds*) aproximadamente iguais. Em cada iteração, um dos *folds* é utilizado como conjunto de validação, enquanto os demais servem para treinamento. O processo é repetido k vezes, garantindo que cada amostra seja utilizada exatamente uma vez para validação. As métricas finais são obtidas pela média dos resultados entre as iterações, reduzindo a variância associada a uma única divisão de treinamento e teste.

No caso específico do REN, a aplicação da validação cruzada demanda cuidados adicionais, devido a, natureza sequencial e contextual das anotações. Divisões aleatórias podem fragmentar documentos ou separar sentenças dependentes, introduzindo enviesamentos e comprometendo a consistência dos resultados. Assim, recomenda-se que o particionamento seja realizado por documento e não por sentença, para preservar as dependências contextuais. Além disso, em *corpora* desbalanceadas, é desejável adotar estratégias de estratificação por tipo ou frequência de entidade, evitando *folds* com distribuições muito distintas de rótulos [Gonzalez Hernandez et al., 2024].

A contaminação de dados (*data leakage*) constitui um dos principais fatores responsáveis pela superestimação do poder preditivo de modelos em PLN. Muitos *corpora* públicos apresentam sobreposição parcial de sentenças, trechos ou padrões linguísticos entre os conjuntos de treinamento e teste, o que pode inflacionar artificialmente as métricas de avaliação. Estudos empíricos demonstram que, na presença desse tipo de sobreposição, o desempenho observado tende a refletir a capacidade do modelo de memorizar padrões já vistos, em vez de sua real habilidade de generalização para dados não observados. Como consequência, recomenda-se a avaliação em conjuntos

externos independentes, bem como a análise da variabilidade dos resultados entre diferentes *folds* de validação cruzada, de modo a obter estimativas mais robustas e confiáveis do desempenho do modelo [Elangovan et al., 2021].

Em trabalhos envolvendo anotações ruidosas ou inconsistentes, variações da validação cruzada têm sido propostas, como a remoção de sentenças repetidas entre *folds* ou a identificação de anotações suspeitas com base em divergências entre iterações [Wang et al., 2019]. Essas estratégias auxiliam na detecção e correção de rótulos de má qualidade, contribuindo para um treinamento mais robusto e confiável.

A validação cruzada desempenha, portanto, um papel essencial na avaliação de modelos de REN, permitindo medições mais realistas do desempenho e fornecendo indícios sobre a estabilidade do modelo em diferentes partições dos dados. Em domínios especializados, como saúde, direito ou segurança pública, nos quais os conjuntos de dados costumam ser pequenos e desbalanceados, sua adoção torna-se ainda mais importante para assegurar a efetividade e a reprodutibilidade dos resultados.

II.3.2 Pseudo-Rotulagem

A pseudo-rotulação é uma técnica central em aprendizado semissupervisionado, na qual um modelo inicialmente treinado em um pequeno conjunto rotulado atribui rótulos a exemplos não rotulados. Esses rótulos gerados automaticamente, chamados de *pseudo-labels*, são incorporados ao processo de treinamento, ampliando o conjunto de dados disponíveis e potencialmente melhorando o desempenho do modelo [Zhu, 2005].

Esse método segue o paradigma de *self-training*, no qual o próprio modelo supervisionado é responsável por expandir o conjunto de dados anotados manualmente e, em seguida, é utilizado para rotular automaticamente textos não anotados, criando pseudo-rótulos de entidades. Embora a pseudo-rotulação já tivesse sido empregada em diferentes contextos, Lee [2013] consolidaram o conceito ao aplicá-lo em redes neurais, propondo um processo iterativo no qual exemplos pseudo-rotulados são adicionados progressivamente ao treinamento, desde que atendam a critérios mínimos de confiança.

Um desafio na pseudo-rotulação para REN é o viés de confirmação: erros na identificação de entidades podem ser reforçados no ciclo de auto-treinamento. Arazo et al. [2020] mostrou que a aplicação de regularizações, como *mixup*, e a seleção de exemplos de alta confiança mitigam esse problema. Pahuja et al. [2021] propôs a integração de dicionários e regras linguísticas para validar automaticamente parte dos pseudo-rótulos antes de incorporá-los ao conjunto de treinamento.

O pseudo-rõtulo em REN, conforme descrito na Figura II.3, é especialmente eficaz quando combinado com técnicas de filtragem por limiar de confiança e *curriculum learning*. Essas estratégias

são vantajosas em domínios com escassez de dados anotados, como medicina, direito ou segurança pública. Nesses casos em que a anotação manual de entidades é cara e exige especialistas, a utilização de métodos automatizados pode ser benéfica. Assim, a pseudo-rotulação se torna uma ferramenta estratégica para ampliar *corpora* de REN e melhorar o desempenho de modelos em cenários reais [Pahuja et al., 2021].

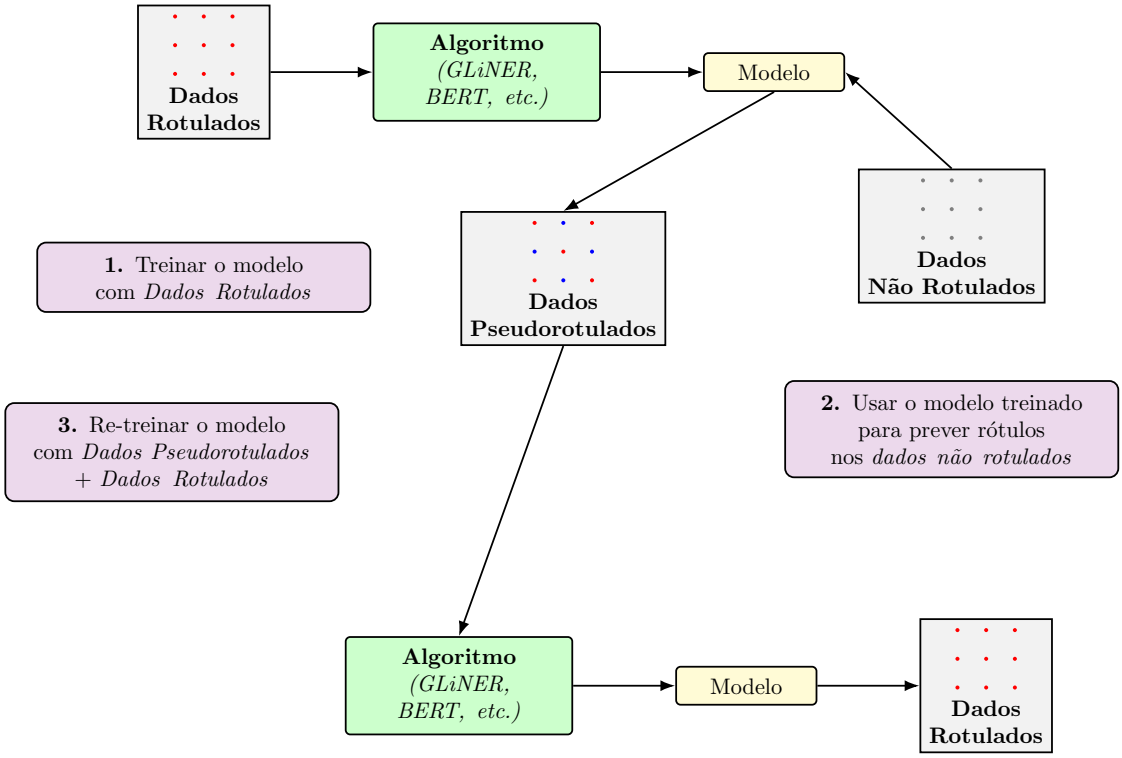


Figura II.3: A figura ilustra o funcionamento básico da técnica de pseudo-rotulação em aprendizado semissupervisionado. Inicialmente (1), um algoritmo é aplicado para treinar o modelo com dados rotulados. Em seguida (2), o modelo treinado prevê rótulos para dados não rotulados, gerando dados pseudo-rotulados. Por fim (3), o modelo é re-treinado utilizando dados rotulados e pseudo-rotulados, buscando melhorar o desempenho geral [Tsang, 2022].

A pseudo-rotulação é especialmente eficaz em domínios onde a anotação manual é custosa ou exige conhecimento especializado, como medicina, direito ou segurança pública. Nesses casos, o uso de pseudo-rótulos permite ampliar substancialmente o conjunto de treinamento a partir de dados não anotados, reduzindo a dependência de especialistas e acelerando o processo de construção do *corpus*.

Apesar de suas vantagens, o uso de pseudo-rotulação requer cuidados metodológicos. A inclusão indiscriminada de exemplos ruidosos pode degradar o desempenho do modelo, especialmente em tarefas sensíveis como o REN, nas quais erros de delimitação ou classificação de entidades podem comprometer fases posteriores do pipeline. Por essa razão, recomenda-se:

- monitorar a qualidade dos pseudo-rótulos por amostragem manual;
- ajustar limiares de confiança de acordo com o desbalanceamento entre entidades;

- avaliar o modelo em validação externa para detectar possíveis amplificações de erro.

Em síntese, a pseudo-rotulação constitui uma abordagem poderosa para ampliar *corpora* e melhorar o desempenho de modelos de REN em cenários com escassez de dados anotados. Quando combinada com estratégias de filtragem, regularização e validação cuidadosa, ela contribui para um aprendizado mais robusto e eficiente em domínios desafiadores.

II.3.3 Calibração de Modelos

A calibração de modelos é um processo estatístico destinado a ajustar as probabilidades previstas por um classificador, de modo que reflitam adequadamente a frequência observada dos eventos reais. Em tarefas de classificação, um modelo bem calibrado é aquele no qual a probabilidade prevista de aproximadamente 70% pertence de fato à classe positiva [Guo et al., 2017]. Esse alinhamento entre previsão e realidade é essencial em aplicações sensíveis, como diagnóstico médico, detecção de fraudes ou análise forense, nas quais as decisões precisam se basear em probabilidades confiáveis [Niculescu-Mizil and Caruana, 2005].

Diversas técnicas de calibração têm sido propostas na literatura. Entre as mais tradicionais, destacam-se:

- **Platt Scaling**: ajusta as probabilidades por meio de uma regressão logística [Platt, 1999];
- **Isotonic Regression**: método não paramétrico que modela a calibração de forma monotônica [Zadrozny and Elkan, 2002];
- **Temperature Scaling**: técnica simples e eficaz para redes neurais profundas, que ajusta uma única temperatura escalar aplicada aos logits do modelo [Guo et al., 2017].

A escolha da técnica depende tanto do tipo de modelo quanto da quantidade de dados disponíveis para ajuste, sendo recomendável o uso de um conjunto de validação independente para evitar *overfitting* durante a calibração.

A fundamentação teórica da calibração está associada às propriedades de modelos probabilísticos bem comportados, que maximizam métricas como *log-loss*, *Negative Log-Likelihood* (NLL) e *Brier score* [Gneiting and Raftery, 2007]. O NLL penaliza fortemente previsões com baixa probabilidade atribuída ao rótulo correto, sendo sensível tanto à precisão quanto à calibração. Já o *Brier score* admite a decomposição clássica de Murphy, separando os componentes de incerteza, resolução e confiabilidade [Murphy, 1973]. A componente de confiabilidade, em particular, relaciona-se diretamente à qualidade da calibração.

Curvas de confiabilidade e calibração

Curvas de confiabilidade (*reliability diagrams*) são ferramentas amplamente utilizadas para diagnosticar a calibração de modelos probabilísticos [Naeini et al., 2015]. Essas curvas particionam o intervalo de probabilidades previstas em faixas (*bins*) e comparam a média das previsões em cada faixa com a fração real de instâncias positivas. Idealmente, os pontos de curva devem se alinhar a diagonal, indicando perfeita calibração.

Dizemos que um classificador está bem calibrado quando, entre todos os casos aos quais o modelo atribui a mesma probabilidade, a fração real de positivos coincide com esse valor. Diferentemente da discriminação (capacidade de separar classes, medida por métricas como AUC e F1), a calibração foca na confiabilidade dos escores como probabilidades [Niculescu-Mizil and Caruana, 2005; Naeini et al., 2015].

Para mensurar desvios de calibração, utilizam-se métricas específicas, entre as quais se destacam:

- **ECE (Expected Calibration Error)**: estima o desvio médio ponderado entre a confiança prevista e a observada;
- **MCE (Maximum Calibration Error)**: mede o maior desvio entre as faixas [Naeini et al., 2015];
- **NLL (Negative Log-Likelihood)**: avalia a qualidade geral da distribuição probabilística prevista.

O NLL apresentado na Equação II.4 mede a qualidade probabilística das predições do modelo, penalizando fortemente casos em que a probabilidade atribuída ao rótulo verdadeiro é baixa. Sua formulação é dada por:

$$NLL = -\frac{1}{N}\sum_{i=1}^{N} \log\, p_\theta(y_i \mid x_i) \tag{II.4}$$

Na equação acima, temos que:

- N – representa o número total de amostras;
- y_i – é o rótulo verdadeiro da amostra i;
- x_i - é a entrada correspondente;
- $p_\theta(y_i \mid x_i)$ é a probabilidade prevista pelo modelo para o rótulo correto.

O *Brier Score* , definido na Equação II.5, avalia diretamente a qualidade probabilística das predições, medindo o erro quadrático entre a probabilidade prevista e o valor observado (0 ou 1). É definido como:

$$BS = \frac{1}{N} \sum_{i=1}^{N} (p_i - y_i)^2 \tag{II.5}$$

Na equação acima, temos que:

- N – é o número de amostras;
- p_i – é a probabilidade prevista para a classe positiva;
- x_i - é a entrada correspondente;
- y_i - é o rótulo verdadeiro.

Diferentemente do NLL, que penaliza mais severamente predições com probabilidades muito baixas para a classe correta, o *Brier Score* avalia o erro quadrático médio das probabilidades, podendo ser decomposto em componentes de incerteza, resolução e confiabilidade. Valores menores indicam predições mais calibradas e, consequentemente, mais confiáveis.

As curvas e métricas de calibração desempenham, portanto, um papel fundamental na avaliação de modelos de PLN, especialmente em cenários com dados desbalanceados ou com entidades raras. Nessas situações, é comum que os modelos superestimem classes majoritárias e subestimem classes minoritárias, comprometendo a confiabilidade das probabilidades previstas e, consequentemente, a qualidade das decisões baseadas nesses modelos.

Calibração no Contexto do REN

No REN, a calibração torna-se ainda mais relevante devido à estrutura multi-rótulo e à natureza assimétrica das entidades. Modelos baseados em *Transformers* frequentemente apresentam excesso de confiança em suas previsões, mesmo quando as probabilidades não correspondem à realidade. Tal problema é agravado em *corpora* desbalanceados, nos quais entidades raras recebem pontuações subestimadas, reduzindo significativamente o *revocação*.

Técnicas como *Temperature Scaling* e *Isolation Regression* mostram-se particularmente úteis nesse cenário por serem simples, eficientes e por não alterarem a ordenação das probabilidades, preservando, assim, a capacidade discriminativa do modelo, ao mesmo tempo em que melhoram a confiabilidade dos escores.

Em pipelines completos de REN, a aplicação de calibração pode:

- aumentar a robustez das decisões baseadas em limiares de confiança;
- melhorar a seleção de spans em técnicas de pseudo-rotulação;

- reduzir vieses decorrentes do desbalanceamento entre classes;
- fornecer probabilidades mais interpretáveis para auditoria e validação.

Assim, a calibração constitui uma etapa essencial no desenvolvimento de modelos modernos de REN, contribuindo para resultados mais confiáveis e consistentes, especialmente em domínios sensíveis e de alta variabilidade.

II.4 Discussão

Este capítulo apresenta os fundamentos teóricos necessários para a compreensão da metodologia e dos experimentos desenvolvidos nos capítulos seguintes. Nesse contexto, os princípios do Processamento de Linguagem Natural estabelecem a base para o tratamento de dados textuais não estruturados, especialmente no que se refere à representação computacional da linguagem e à modelagem das informações contextuais presentes nos relatos criminais do Disque Denúncia.

A Seção II.1, dedicada ao Reconhecimento de Entidades Nomeadas, fornece o arcabouço conceitual que sustenta a principal tarefa investigada neste trabalho: a identificação automática de entidades do tipo Pessoa, Local e Organização em textos informais. Esses fundamentos orientam tanto o processo de anotação do *corpus* quanto a definição das estratégias de modelagem adotadas posteriormente.

As métricas de avaliação apresentadas na Seção II.1.3 são essenciais para a análise dos resultados experimentais, permitindo comparar diferentes configurações de modelos e quantificar, de forma objetiva, o desempenho das abordagens propostas.

No que se refere às técnicas de modelagem, a Seção II.2.3 introduz o modelo GLiNER, cuja escolha se justifica por sua capacidade de generalização em cenários de vocabulário aberto, característica particularmente relevante para textos informais. A Seção II.2.4 complementa essa discussão ao apresentar o ajuste fino, mecanismo fundamental para adaptar modelos pré-treinados ao domínio específico dos relatos criminais.

A Seção II.3.2 apresenta a técnica de pseudo-rotulação, que desempenha um papel central na metodologia proposta ao permitir a expansão do conjunto de treinamento em um cenário com escassez de dados anotados. Sua inclusão é crucial para compreender a estratégia semissupervisionada adotada neste trabalho e seu impacto no desempenho dos modelos.

As técnicas de validação cruzada (Seção II.3.1) e calibração de modelos (Seção II.3.3) são apresentadas como mecanismos complementares para garantir a robustez e a confiabilidade dos resultados experimentais. Enquanto a validação cruzada permite estimar de forma mais estável a capacidade de generalização dos modelos, a calibração assegura que as probabilidades previstas sejam interpretáveis e consistentes com a frequência observada dos eventos.

Capítulo III Trabalhos Relacionados

Este capítulo está organizado em quatro seções principais. Na Seção III.1, são detalhadas as etapas do processo de busca e seleção das publicações. A Seção III.2 apresenta os critérios utilizados para determinar a elegibilidade e a priorização dos estudos analisados. Em seguida, na Seção III.3, são apresentados os resultados da revisão, com destaque para os principais trabalhos selecionados e sua relação com a abordagem proposta nesta dissertação. Por fim, a Seção III.4 traz a conclusão da análise, com base nos resultados obtidos ao longo do processo de leitura e avaliação dos estudos.

III.1 Seleção de publicações

Para a realização desta revisão, foram selecionadas as bases de dados Instituto de Engenheiros Eletricistas e Eletrônicos (IEEE), *Association for Computational Linguistics* (ACL) *Anthology* e Scopus, escolhidas em razão da credibilidade, abrangência e relevância de seu acervo disponível para pesquisas em ciência da computação e inteligência artificial. O escopo dos documentos incluídos restringiu-se a artigos publicados em revistas científicas e trabalhos apresentados em conferências, assegurando a qualidade e o rigor metodológico das publicações analisadas.

Foi estabelecido um recorte temporal compreendendo os últimos oito anos (2017 a 2025), alinhado à evolução recente dos modelos e técnicas de PLN, em especial no que se refere ao REN. Para a construção da estratégia de busca, os termos foram organizados em três grupos principais: (i) reconhecimento de entidades nomeadas, (ii) linguagem informal e contextos relacionados e (iii) segurança pública, violência e aprendizado semissupervisionado. As combinações utilizadas foram as seguintes:

- **Reconhecimento de Entidades Nomeadas**
  - "Named Entity Recognition"
  - "Entity Recognition"
  - "NER"
- **Linguagem Informal e Contextos Relacionados**
  - "Informal Language"

  - "Hate Language"
  - "Hate Speech"
  - "Slang"
  - "Internet Language"
  - "Colloquial Language"
  - "Social Media"

- **Segurança Pública e Violência**
  - "Public Security"
  - "Public Safety"
  - "Crime"
  - "Criminal"
  - "Violence"
  - "Violent"
  - "Hate Crime"
  - "Law Enforcement"
  - "Abusive Language"

- **Aprendizado Semissupervisionado e Pseudo-Rotulação**
  - "Pseudo-Label"
  - "Pseudolabeling"
  - "Semi-supervised"
  - "Weak Supervision"
  - "Distant Supervision"
  - "Self-training"
  - "Bootstrapping"
  - "Label Propagation"
  - "Noisy Labels"
  - "Silver Data"

A partir desses grupos, foram realizadas combinações utilizando o operador lógico **AND** entre os diferentes conjuntos temáticos, permitindo recuperar estudos relacionados ao reconhecimento de

entidades nomeadas em contextos de linguagem informal, segurança pública, violência e técnicas de aprendizado semissupervisionado.

A seleção dos artigos foi conduzida com base nas diretrizes do método *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) [Page et al., 2021], as quais estruturam o processo de revisão em diferentes etapas, abrangendo identificação, triagem, avaliação de elegibilidade e inclusão final dos estudos. A Figura III.1 apresenta o fluxo completo desse processo, incluindo o número de registros encontrados, os motivos de exclusão e o total de publicações selecionadas.

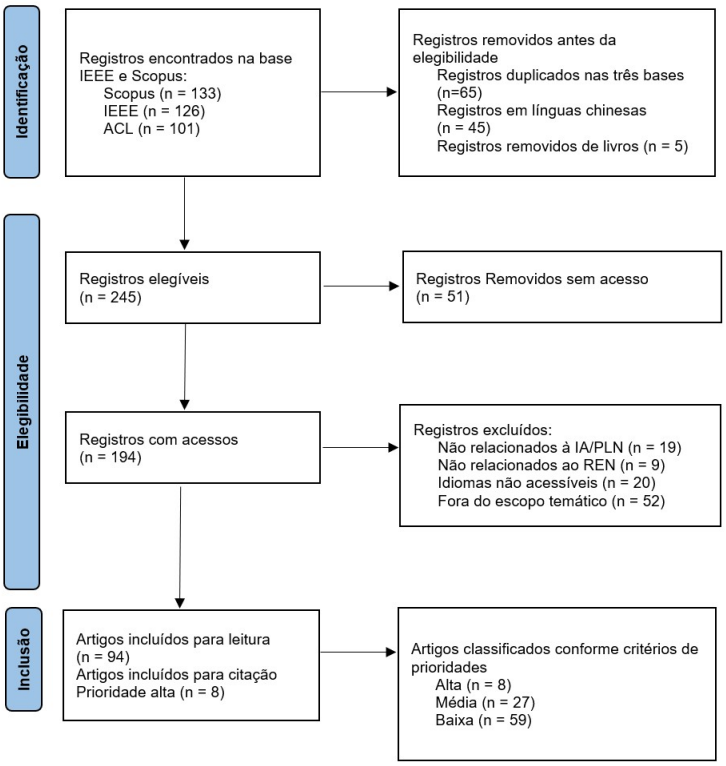


Figura III.1: Fluxo de seleção de artigos com base na metodologia PRISMA, detalhando as etapas de identificação das bases, remoção de registros, avaliação de elegibilidade, exclusões por critérios definidos e classificação final dos artigos incluídos.

Conforme a Figura III.1, o processo é dividido em três etapas principais: identificação, elegibilidade e inclusão. Na etapa de identificação, foram encontrados registros nas bases Scopus, ACL Anthology e IEEE. Na fase de elegibilidade, foram aplicados critérios de exclusão, resultando em um subconjunto de artigos considerados relevantes. Por fim, na etapa de inclusão, os artigos foram classificados com base em prioridades para citação e análise.

III.2 Critérios de inclusão, exclusão e priorização

A aplicação da metodologia PRISMA exigiu a definição de um conjunto claro de critérios específicos de inclusão, exclusão e priorização dos estudos mais relevantes ao escopo desta dissertação.

Esses critérios orientaram a triagem e a avaliação de elegibilidade, garantindo a consistência e o rigor metodológico da revisão.

III.2.1 Critérios de Inclusão

Os critérios de inclusão foram estabelecidos para assegurar que as publicações selecionadas estivessem alinhadas ao objeto de pesquisa, contemplando aspectos essenciais à REN e ao domínio da segurança pública. Assim, foram considerados elegíveis os estudos que atendessem aos seguintes requisitos:

- **Critério 1** - Estudos que abordam técnicas, modelos ou metodologias relacionadas ao Reconhecimento de Entidade Nomeada
- **Critério 2** - Publicações que utilizam bases de dados relacionadas à violência, discurso de ódio, segurança pública ou linguagem informal associada a esses contextos.
- **Critério 3** - Artigos publicados entre 2021 e 2025, período marcado por avanços significativos em modelos de PLN.

III.2.2 Critérios de Exclusão

Os critérios de exclusão foram definidos para remover estudos que não atendiam aos objetivos da revisão ou que poderiam comprometer a coesão temática dos resultados. Foram excluídas as publicações que:

- **Critério 1** – Não apresentavam relação com Inteligência Artificial (IA), PLN ou tarefas de extração de informação
- **Critério 2** – Não abordavam técnicas de identificação, extração ou classificação de entidades relevantes ao REN
- **Critério 3** – Estavam redigidas em idiomas cuja compreensão ou validação não era possível no contexto da pesquisa (como o chinês ou coreano), impossibilitando a avaliação detalhada dos métodos
- **Critério 4** – Trabalhos cujo objeto de estudo não possui relação com o corpus analisado nesta pesquisa, especialmente no que se refere ao contexto de violência, discurso de ódio ou segurança pública.

III.2.3 Critérios de Priorização

Após a triagem e a avaliação de elegibilidade, os artigos selecionados foram classificados conforme sua relevância para os objetivos da dissertação. A definição de prioridades auxiliou na organização da análise e na identificação dos estudos com maior impacto potencial.

- **Prioridade alta** - Trabalhos que apresentam técnicas de REN aplicadas a bases de dados de violência, discurso de ódio ou linguagem informal, com foco em metodologias inovadoras ou modelos avançados, como arquiteturas baseadas em LLM.
- **Prioridade média** - Estudos que tratam da identificação de entidades de forma indireta, como por meio da análise de tópicos, classificação de temas ou tarefas auxiliares relacionadas.
- **Prioridade baixa** - Publicações que utilizam abordagens baseadas em imagens, grafos ou técnicas distantes do escopo textual central da pesquisa.

III.3 Resultados

A metodologia PRISMA, conforme ilustrada na Figura III.1, é aplicada para estruturar de forma sistemática o processo de identificação, triagem e seleção de publicações relacionadas ao REN em contextos de violência, discurso de ódio e linguagem informal. A revisão foi concluída em 20/03/2025, revisitada em 24/01/2026 e foi feita a aplicação dos critérios de inclusão, exclusão e priorização.

Durante a fase de identificação inicial, um conjunto abrangente de artigos foi recuperado a partir das combinações de termos relacionados ao REN, segurança pública, violência e linguagem informal. Em seguida, na triagem e na avaliação de elegibilidade, os registros foram submetidos aos critérios previamente definidos, resultando na remoção dos estudos que não atendiam aos requisitos temáticos, metodológicos ou linguísticos. Após essa filtragem, permaneceu um conjunto reduzido, porém representativo, de artigos alinhados aos objetivos desta dissertação.

III.3.1 Publicações selecionadas

O estudo conduzido por da Silva et al. [2023] tem como objetivo investigar os resultados obtidos a partir de bases de dados oriundas de redes sociais globais que contêm manifestações de discurso de ódio, por meio da aplicação do modelo BERTimbau em quatro tarefas distintas: análise de sentimentos, extração de aspectos, detecção de discurso de ódio e detecção de ironia. Para a realização dos experimentos, os autores utilizam conjuntos de dados em língua portuguesa, cada um correspondente a uma tarefa específica. Os textos foram extraídos de diferentes plataformas, apresentando características como polaridades sentimentais distintas, linguagem ofensiva e presença

de ironia. Duas variantes do modelo BERTimbau (base e large) são submetidas a ajuste fino, com ajuste de hiperparâmetros e aplicação de técnicas de pré-processamento, incluindo *oversampling* e *data augmentation*, com o objetivo de mitigar o desbalanceamento entre classes. Os resultados evidenciaram melhorias significativas em comparação com as abordagens tradicionais utilizadas anteriormente.

Gomes Jr. et al. [2023] apresentam um *framework* voltado à coleta, unificação e enriquecimento de dados sobre pessoas desaparecidas no Brasil, integrando informações de diversas fontes na web, como Organizações Não Governamentais (ONGs) e órgãos governamentais. O objetivo dos autores é melhorar a qualidade desses registros e oferecer subsídios consistentes para análises e para a tomada de decisões relacionadas ao fenômeno dos desaparecimentos. Para a composição dos *corpora*, foram coletados dados de plataformas mantidas por entidades civis e governamentais, contendo informações como gênero, idade, data e local de desaparecimento, além de detalhes fornecidos por familiares. Os registros apresentam desafios significativos, como a ausência de variáveis importantes (por exemplo, etnia), inconsistências textuais e falta de padronização entre as instituições. Para superar essas limitações, foi proposto um *framework* baseado em técnicas de mineração de dados e enriquecimento semântico, capaz de estruturar e padronizar as informações. A extração de dados é realizada por meio de *scraping*, utilizando consultas XPath e expressões regulares adaptadas a cada site. Em seguida, técnicas de classificação, como árvores de decisão, regressão logística e *random forest*, foram aplicadas para organizar os registros e identificar padrões relevantes, incluindo faixa etária, características físicas e iniciativas estaduais. Os resultados demonstraram que a abordagem contribui para o aprimoramento das bases públicas, ampliando sua consistência e utilidade em análises sociais e institucionais.

O trabalho de Ferreira et al. [2022] propõe uma metodologia automatizada para extração de estatísticas relacionadas a eventos de segurança pública a partir de microtextos publicados no Twitter sobre o estado do Rio de Janeiro. A abordagem utiliza técnicas de PLN e REN para identificar eventos criminais, locais, datas e horários em textos informais marcados por gírias e regionalismos. O modelo é desenvolvido com aprendizado supervisionado, utilizando exemplos anotados manualmente e técnicas baseadas em *word embeddings*, *Bloom embeddings* e redes neurais profundas com conexões residuais. Os resultados demonstram desempenho superior a 95% no reconhecimento das entidades relacionadas aos eventos de segurança pública, evidenciando o potencial da mineração de dados e do PLN aplicados a contextos de violência urbana.

Zhu and Bhat [2021] aborda o problema da detecção automática de frases eufemísticas, frequentemente utilizadas em redes sociais para ocultar termos relacionados ao uso de drogas. O *corpus* utilizado consiste em textos provenientes de fóruns online contendo postagens sobre drogas. A partir desse *corpus*, foi extraída uma lista de palavras e frases candidatas por meio de técnicas de

mineração de frases. Para esta etapa, é empregado o AutoPhrase, responsável pela extração das expressões mais relevantes do conjunto textual. Em seguida, o modelo Word2Vec é utilizado para calcular as similaridades entre os *embeddings* das frases e das palavras-chave, filtrando os candidatos irrelevantes. O modelo SpanBERT é ajustado para prever a probabilidade de que determinadas frases funcionassem como substitutas de palavras-chave em sentenças mascaradas. Esse método apresentou desempenho superior a *baselines* como Word2Vec e EigenEuph, alcançando uma melhoria entre 20% e 50% na detecção de eufemismos. A avaliação é realizada utilizando a métrica *Precision at k* (P@k), com valores reportados para P@10, P@20, P@30 e P@50, permitindo a comparação do desempenho em diferentes recortes do ranking. Os resultados demonstraram que o modelo *Euphemistic Phrase Detection* (EPD) superou significativamente os métodos comparados.

Dharviyanti and Wilantika [2024] propõem um método para a extração de informações relacionadas a crimes a partir de notícias online publicadas na Indonésia, buscando suprir limitações das estatísticas criminais tradicionais, que apresentam muitas vezes dados desatualizados ou insuficientemente detalhados. O *corpus* analisado é composto por 53.336 notícias publicadas entre janeiro de 2021 e dezembro de 2022, selecionadas com base em palavras-chave representativas de categorias criminais presentes no código penal do país. Para identificar entidades relevantes, como datas, tipos de crimes, locais e pessoas, foi empregado um modelo de REN baseado em regras. Técnicas como expressões regulares foram utilizadas para o reconhecimento de padrões de data, enquanto os componentes *Token Matcher* e *Phrase Matcher* da biblioteca Spacy são aplicados para a identificação de tipos criminais e localizações. A avaliação, conduzida por meio de rotulagem manual de amostras aleatórias, resultou em uma pontuação média de 90,71% na métrica *F1*, além de valores satisfatórios de precisão e *revocação*.

Rothe et al. [2023] investigaram a detecção e a identificação de palavras de gíria em textos, com foco na compreensão da linguagem informal presente em redes sociais e na ampliação das capacidades de sistemas de Processamento de Linguagem Natural. O *corpus* utilizado é o Slangvolution Twitter Data, disponível no Kaggle, composto por mais de 225.000 *tweets* rotulados de acordo com as gírias presentes nas mensagens. O pré-processamento inclui a remoção de pontuação, a normalização para minúsculas e a vetorização textual por meio do CountVectorizer e *Term Frequency-Inverse Document Frequency* (TF-IDF). Para a tarefa de detecção, são utilizados modelos como *Gaussian Naive Bayes* e *Random Forest*, enquanto as arquiteturas *Long Short-Term Memory* (LSTM) e *Bidirectional recurrent neural networks* (Bi-LSTM) foram aplicadas para capturar dependências contextuais em textos curtos. Para a identificação de gírias em nível de entidade, o modelo BERT é utilizado com o esquema BILOU, adaptado ao domínio da linguagem informal. Os resultados foram expressivos: o classificador *Random Forest* atingiu uma acurácia de 96,01% na detecção de gírias, enquanto os modelos baseados em BERT apresentaram um desempenho sólido

na identificação de entidades rotuladas como expressões informais.

O trabalho de Caselli et al. [2021] propõe o HateBERT, um modelo de linguagem baseado em BERT que passa por uma etapa adicional de pré-treinamento voltada especificamente à detecção de linguagem abusiva em inglês. Diferentemente de abordagens que utilizam modelos genéricos ajustados apenas por ajuste fino, os autores realizam o re-treinamento do BERT base-uncased utilizando o conjunto de dados RAL-E, composto por aproximadamente 1,5 milhão de comentários extraídos de comunidades do Reddit banidas por promoverem discurso ofensivo, abusivo ou de ódio. O modelo é avaliado em três tarefas distintas, detecção de linguagem ofensiva, linguagem abusiva e discurso de ódio, utilizando os conjuntos OffensEval 2019, AbusEval e HatEval. Os resultados demonstram que o HateBERT supera consistentemente o BERT genérico em termos de F1, tanto em avaliações intra-dataset quanto em experimentos de portabilidade entre bases diferentes. O estudo evidencia que o pré-treinamento em domínios específicos contribui para representações mais robustas e sensíveis às nuances da linguagem abusiva, destacando a relevância de modelos especializados para aplicações em contextos de violência e discurso de ódio em mídias sociais.

Otal et al. [2024] propõe uma arquitetura baseada em LLM para apoio à gestão de crises e resposta a emergências, explorando dois cenários principais: o suporte ao despacho de chamadas de emergência e a colaboração pública por meio da análise de mensagens em redes sociais. Os autores utilizam modelos de linguagem de grande escala, como LLaMA2 e Mistral, ajustados por meio de ajuste fino supervisionado, para classificar mensagens emergenciais para extrair informações relevantes e fornecer orientações contextuais em tempo real. O estudo utiliza corpora de violência ou discurso de ódio, sua relevância reside na demonstração do potencial dos LLMs para lidar com linguagem informal, dados ruidosos e comunicações sensíveis no domínio da segurança pública. Dessa forma, o trabalho contribui de forma indireta para o entendimento de como arquiteturas modernas de PLN podem ser aplicadas em contextos críticos, complementando as abordagens analisadas nesta dissertação

III.3.2 Comparação

A Tabela III.1 apresenta uma síntese das principais características dos estudos discutidos nesta seção, permitindo uma visão comparativa das técnicas empregadas, dos modelos utilizados e das particularidades de cada abordagem. A tabela evidencia elementos como o uso ou não de modelos baseados em BERT, a presença de tarefas relacionadas ao reconhecimento de entidades nomeadas, a aplicação em bases de dados contendo informações sobre violência ou discurso de ódio, e o tratamento da linguagem informal.

A comparação entre os trabalhos evidencia que, embora existam esforços relevantes em diferentes domínios e idiomas, o uso de técnicas de reconhecimento de entidades nomeadas aplicadas especifi-

camente a bases de dados de violência em língua portuguesa ainda é pouco explorado. Observa-se também uma variação significativa nas metodologias adotadas, que vão desde arquiteturas tradicionais, como *Bidirectional Long Short-Term Memory* (BiLSTM), combinadas a camadas *Conditional Random Field* (CRF), até modelos mais robustos baseados em BERT e outras variantes de LLM. Além disso, alguns trabalhos focam em desafios linguísticos específicos, como a detecção de eufemismos ou gírias em redes sociais, enquanto outros priorizam a padronização e o enriquecimento de dados em contextos sensíveis, como o desaparecimento de pessoas ou a cibersegurança.

Essa diversidade metodológica permite compreender como diferentes abordagens tratam problemas associados à informalidade textual, à variabilidade linguística e à presença de ruído nos dados, fornecendo subsídios importantes para contextualizar a relevância e o posicionamento da proposta apresentada nesta dissertação.

Tabela III.1: Visão geral das características dos trabalhos relacionados.

Artigo	Técnica Principal	Idioma	Considera BERT	Reconhecimento de Entidades	Base de Dados Violência	Considera Linguagem Informal	Pseudo Rotulação
da Silva et al. [2023]	Análise de Sentimento e Extração de Aspectos em base de dados de ódio	PT-BR	✓	✓	✓	✓	-
Gomes Jr. et al. [2023]	Classificação de Entidades para características de pessoa desaparecidas	PT-BR	-	✓	✓	-	-
Ferreira et al. [2022]	Extração automatizada de estatísticas de eventos de segurança pública relatados por meio de microtextos em redes sociais.	PT-BR	-	✓	✓	✓	-
Zhu and Bhat [2021]	Detecção de eufemismo e comparação dos modelos	EN	✓	✓	✓	✓	-
Dharviyanti and Wilantika [2024]	Classificação de entidades de pessoas para as bases de dados de notícias sobre violência	EN	✓	✓	✓	-	-
Rothe et al. [2023]	Identificação de gírias utilizando BERT e comparação de modelos	EN	✓	✓	-	✓	-
Caselli et al. [2021]	Identificação de linguagens abusivas utilizando BERT	EN	✓	✓	-	✓	-
Otal et al. [2024]	Classificação e análise de linguagem informal em contextos de emergência com LLMs	EN	-	✓	✓	✓	-
Esta dissertação	Reconhecimento de entidades nomeadas em relatos criminais informais com apoio de metadados	EN	✓	✓	✓	✓	✓

III.4 Discussão

Foi realizada uma análise abrangente das publicações disponíveis nas principais bases científicas, como Scopus, IEEE e ACL *Anthology*, contemplando o período de 2017 a 2025. Esse recorte temporal coincide com a consolidação e a expansão dos modelos de PLN. A partir desta revisão, é possível observar a evolução das técnicas aplicadas ao REN, bem como a adaptação dos modelos a diferentes tipos de *corpus*, variando entre textos formais, informais, sensíveis ou provenientes de redes sociais.

Os artigos publicados entre 2017 e 2021 indicam uma predominância de estudos que associam o REN à análise de sentimentos e a modelos menos robustos, em um momento em que as técnicas de aprendizado profundo ainda estavam sendo amplamente exploradas. No entanto, a partir de 2022, nota-se uma mudança significativa. O avanço e a popularização de modelos derivados do BERT e de outras arquiteturas pré-treinadas ampliaram a capacidade de captura de nuances semânticas e contextuais, permitindo abordagens mais precisas e eficientes em tarefas de extração de entidades.

Apesar da diversidade de estudos identificados, verificou-se que há poucas pesquisas dedicadas ao reconhecimento de entidades nomeadas em bases de dados sobre violência na língua portuguesa. Em contrapartida, há produções mais expressivas em outros idiomas, evidenciando a necessidade de aprofundamento do tema no contexto brasileiro. A presente dissertação está alinhada justamente a essa lacuna, ao propor o uso de técnicas modernas de processamento de linguagem natural para tratar a linguagem coloquial em português, característica predominante nos relatos recebidos pelo Disque Denúncia, *corpus* adotado para esta pesquisa.

Assim, a revisão realizada não apenas contextualiza o estado da arte sobre o tema, mas também reforça a relevância e a originalidade da proposta desta dissertação ao demonstrar tanto o avanço metodológico alcançado pela comunidade científica quanto as limitações existentes, que justificam o desenvolvimento da abordagem apresentada neste trabalho.

Capítulo IV Reconhecimento de Entidades Nomeadas em Relatos Criminais

Neste capítulo, apresenta-se a metodologia proposta para enfrentar o problema de REN em relatos criminais redigidos em linguagem informal, com ênfase na utilização de metadados estruturados como fonte adicional de contexto.

A abordagem metodológica contempla o uso combinado de aprendizado supervisionado e semissupervisionado. Inicialmente, um modelo de linguagem de grande porte é refinado por meio de ajuste fino com base em um *corpus* manualmente anotado, contendo relatos oriundos do aplicativo do Disque Denúncia. Em seguida, esses modelos são utilizados para gerar pseudo-anotações sobre um segundo conjunto de relatos, significativamente maior e ainda não rotulado.

Adicionalmente, a metodologia proposta explora a informação disponível em metadados associados a cada relato, como bairro, logradouro, tipo de ocorrência e data, a fim de enriquecer a representação dos textos e melhorar a capacidade de generalização dos modelos de REN. Esses metadados são incorporados ao *pipeline* tanto durante o pré-processamento quanto na construção de estratégias de entrada textual e validação cruzada.

As seções seguintes detalham a formulação do problema, as características dos dados utilizados, a preparação dos conjuntos de treino e validação, bem como os procedimentos de treinamento, pseudo-rotulagem e avaliação dos modelos desenvolvidos. A Seção IV.1 trata da formulação do problema analisado. Na Seção IV.2, são descritas as principais características dos relatos criminais utilizados neste estudo e apresentados os dados empregados na pesquisa. Na Seção IV.3, detalha-se cada etapa da metodologia proposta.

IV.1 Formulação do Problema

Este trabalho investiga o problema de REN em relatos criminais informais provenientes do Disque Denúncia. Diferentemente de corpora tradicionalmente utilizados em PLN, esses relatos apresentam linguagem não padronizada, com erros ortográficos, abreviações e estruturas sintáticas irregulares, o que compromete o desempenho de modelos treinados em domínios formais.

Além disso, o cenário é caracterizado pela escassez de dados anotados manualmente, o que limita a aplicação direta de abordagens supervisionadas convencionais. Ao mesmo tempo, cada relato é acompanhado por metadados estruturados, como tipo de ocorrência, localização e data.

Esse metadados fornecem informações contextuais potencialmente úteis, mas que são, em geral, pouco exploradas em abordagens de REN.

Nesse contexto, o problema de pesquisa pode ser formulado como:

> *Como reconhecer automaticamente entidades nomeadas (pessoas, locais e organizações) em relatos criminais escritos em linguagem informal, explorando de forma integrada o conteúdo textual e os metadados estruturados disponíveis, em um cenário com escassez de dados rotulados?*

A investigação desse problema envolve os seguintes desafios principais:

- Modelar adequadamente a variabilidade e o ruído presentes na linguagem informal dos relatos;
- Mitigar a limitação de dados anotados, explorando estratégias semissupervisionadas;
- Integrar metadados heterogêneos ao processo de modelagem de forma eficaz;
- Avaliar o desempenho de arquiteturas modernas de REN, como o GLiNER, nesse domínio específico.

Dessa forma, este trabalho busca avançar o estado da arte ao propor uma abordagem que combina modelos de linguagem ajustados ao domínio, uso de pseudo-rotulagem e incorporação de metadados, visando melhorar a extração de informações estruturadas a partir de relatos criminais informais descritos na língua portuguesa.

IV.2 *Corpora* de Relatos Criminais

Esta pesquisa utiliza dados provenientes de relatos criminais registrados por meio do aplicativo Disque-Denúncia. Tais relatos consistem predominantemente no conteúdo textual redigido pelos próprios denunciantes, contendo descrições das ocorrências. No aplicativo, o usuário seleciona a categoria principal do relato, o que auxilia na organização e priorização do atendimento pelas forças policiais.

Utilizam-se dois *corpora* extraídos do aplicativo Disque Denúncia, os quais desempenham papéis complementares na metodologia proposta. O primeiro, denominado DD_{small}, corresponde a um *corpus* manualmente anotado, composto por relatos criminais registrados entre 10/04/2016 e 12/09/2017. Além do texto descritivo da ocorrência, cada registro contém entidades nomeadas dos tipos `Pessoa`, `Local` e `Organização`, identificadas segundo o esquema BIO, bem como metadados estruturados, como categoria da ocorrência, bairro e logradouro.

O processo de anotação manual foi conduzido pelos anotadores humanos, seguindo diretrizes previamente definidas para a identificação consistente das entidades nomeadas. Posteriormente, as

anotações foram revisadas e validadas pelo autor desta dissertação, com o objetivo de assegurar a qualidade e a uniformidade das marcações realizadas.

Embora apresente dimensão relativamente reduzida, esse conjunto constitui a principal fonte de dados supervisionados para o treinamento, validação e avaliação dos modelos desenvolvidos.

O segundo conjunto de dados, denominado DD_{large}, é composto por relatos criminais registrados entre 30/07/2016 e 07/12/2022. Diferentemente do DD_{small}, esse *corpus* não possui anotações manuais de REN, mas contém os mesmos tipos de metadados estruturados e um volume significativamente maior de registros.

Em razão dessas características, o DD_{large} é empregado nas etapas de aprendizado semissupervisionado, servindo como base para a geração de pseudo-rótulos a partir de modelos previamente treinados no conjunto anotado.

Ressalta-se que não foi realizado qualquer procedimento de deduplicação entre os conjuntos DD_{small} e DD_{large}. Consequentemente, relatos presentes no conjunto anotado podem também estar contidos no *corpus* de maior escala, uma vez que ambos foram extraídos de períodos parcialmente sobrepostos da base histórica do Disque Denúncia.

A utilização conjunta desses dois *corpora* possibilita combinar a qualidade das anotações manuais disponíveis no DD_{small} com a ampla cobertura de dados do DD_{large}. Dessa forma, torna-se viável ampliar o conjunto de treinamento por meio de pseudo-rotulação, explorando simultaneamente informações textuais e metadados estruturados para aprimorar o desempenho dos modelos de reconhecimento de entidades nomeadas. A Tabela IV.1 apresenta uma visão geral comparativa das principais características de ambos os conjuntos de dados.

Tabela IV.1: Resumo dos *corpora* utilizados no estudo

Característica	DD_{small}	DD_{large}
Número de relatos	5.230	368.854
Média de tokens por relato	91	82
Total de entidades anotadas	20.101	–
Utilização de metadados	Não	Sim
Utilização no experimento	Treino/validação	Ajuste fino

Em ambos os *corpora*, são encontrados relatos pertencentes a categorias que, segundo da Silva and Melo [2020], tendem a receber maior prioridade no processo de triagem e resposta das forças policiais. As categorias selecionadas são: Armas, Tráfico de Drogas e Armas, Roubos de Carga e Veículos, Roubos em Geral e Homicídios.

Além do campo destinado aos relatos criminais, ambos os *corpora* contêm diversos campos preenchidos com os metadados. Entre esses campos, destacam-se aqueles apresentados na Tabela IV.2.

Tabela IV.2: Campos de metadados existentes em cada relato.

Campo de Metadado	Descrição
id	Campo para identificação numérica.
numEnvolvidos	Quantidade de envolvidos no relato criminal.
Categoria	Categoria do relato criminal.
cidadeLocal	Cidade da ocorrência.
logradouroLocal	Logradouro do local da ocorrência.
numeroLocal	Número do local da ocorrência.
complementoLocal	Complemento do local da ocorrência.
bairroLocal	Bairro do local da ocorrência.
pontodeReferenciaLocal	Ponto de referência do local da ocorrência.
data	Data e hora da ocorrência criminal.
lida	Indica se a ocorrência foi lida pelas forças policiais.

Os relatos em cada corpus estão distribuídos pelas categorias apresentadas na Tabela IV.3.

Tabela IV.3: Relatos criminais por categorias do DD_{small} e DD_{large}

Categoria	DD_{small}	DD_{large}
Armas	379	362
Homicídio	380	12.552
Roubos, Cargas e Veículo	900	901
Roubos em Geral	332	73.728
Tráfico de Drogas/Armas	3.239	281.311

A Figura IV.1 apresenta a distribuição percentual dos relatos criminais nos *corpora* DD_{small} e DD_{large} conforme faixas de tamanho, categorizadas com base no número de palavras por relato. A representação em gráficos de barras verticais permite observar a predominância de relatos de menor extensão em ambos os conjuntos, evidenciando padrões relevantes para análises linguísticas e estatísticas. Além disso, a comparação entre os dois *corpora* destaca diferenças na concentração das faixas de tamanho, refletindo variações na extensão dos relatos.

Neste estudo, o processo de reconhecimento de entidades nomeadas concentra-se em três categorias principais de entidades: **Pessoa**, **Local** e **Organização**. A categoria **Pessoa** refere-se a indivíduos mencionados nos relatos criminais, incluindo suspeitos, vítimas ou outros participantes citados no texto. A categoria **Local** abrange referências geográficas ou espaciais relevantes para a ocorrência, como bairros, ruas, comunidades e estabelecimentos. Já a categoria **Organização** corresponde a instituições formais ou informais mencionadas nos relatos, como forças policiais, facções

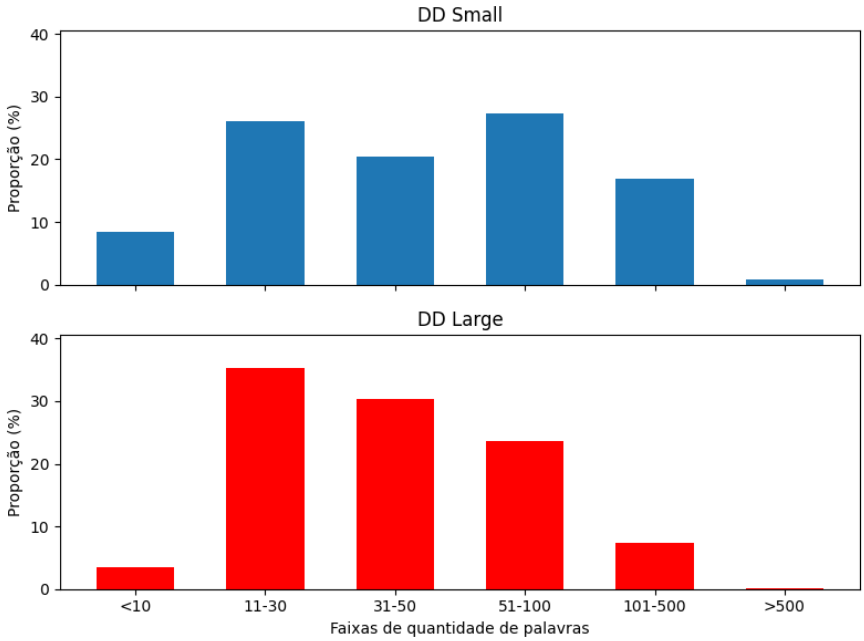


Figura IV.1: Proporção de palavras por relato nos *corpora* DD_{small} e DD_{large}, considerando diferentes faixas de quantidade de palavras. Os gráficos são apresentados separadamente e utilizam a mesma escala no eixo vertical, permitindo comparação direta entre as distribuições do tamanho dos relatos.

criminosas ou órgãos públicos. Essas categorias são definidas considerando a natureza dos relatos criminais e sua relevância para o apoio à análise de informações em segurança pública.

Para a identificação das entidades de organizações, localizações e pessoas, são empregados os modelos previamente treinados por ajuste fino, complementados pela pseudo-rotulagem para ampliar o conjunto de dados. Essa estratégia enriquece o *corpus* com novas entidades geradas automaticamente, com base nas predições de um modelo ajustado a partir do conjunto manualmente anotado.

A Tabela IV.4 apresenta uma visão comparativa das principais estatísticas descritivas dos corpora DD_{small} e DD_{large}, evidenciando tanto a diferença de escala entre os conjuntos quanto suas características textuais básicas.

Os metadados presentes no DD_{large}, preenchidos no momento da denúncia: como `Bairro`, `Cidade`, `Ponto de referência` e `Logradouro` apresentam níveis distintos de completude. A Tabela IV.5 descreve suas porcentagens de preenchimento e cardinalidade, considerando os 368.854 relatos da base.

A Tabela IV.6 apresenta os dez valores mais recorrentes para determinados campos de metadados no $DD_{\textbf{large}}$, evidenciando sua distribuição e importância para análises futuras.

Tabela IV.4: Estatísticas gerais dos corpora DD_{small} e DD_{large}.

Descrição	DD_{small}	DD_{large}
Número de relatos	5.230	368.854
Quantidade de palavras	360.582	18.598.376
Quantidade de Termos	20.101	-
Média de palavras por relatos criminais	68.94	50,70
Mediana de palavras por relatos criminais	45	37
Mínimo de palavras por relatos criminais	1	1
Máximo de palavras por relatos criminais	2.029	5.313
Tamanho total (em MB)	3	178

Tabela IV.5: Estatísticas da completude e cardinalidade dos metadados presentes no *corpus* $DD_{\textbf{large}}$.

Metadado	Porcentagem (%)	Categorias Únicas
`Bairro`	92,82	8.695
`Logradouro`	86,18	42.694
`Ponto de Referência`	87,42	64.446
`Cidade`	92,87	2.502

Tabela IV.6: Dez valores mais recorrentes dos Metadados do $DD_{\mathbf{large}}$

Bairro	Qtd.	**Logradouro**	Qtd.	**Ponto de Referência**	Qtd.	**Cidade**	Qtd.
Coréia	23.518	**Becos**	5.813	**Chatuba**	10.429	**Mesquita**	57.113
Chatuba	17.175	**Vielas**	5.301	**Caixa d'Água**	1.563	**Rio de Janeiro**	25.318
Juscelino	6.114	**Rua Valinhos**	1.713	**Escadão**	1.535	**RJ**	23.395
Centro	5.872	**Rua Caruaru**	1.432	**Próximo a padaria verão vermelho na estrada adhema bebiano**	1.505	**Nova Iguaçu**	10.015
Alto Uruguai	3.955	**Rua Carioca**	1.428	**Alto Uruguai**	1.314	**São Gonçalo**	9.364
Engenho da Rainha	3.177	**Avenida Brasil**	1.359	**Morro Coréia**	932	**Nilópolis**	5.420
k11	2.886	**Rua da Vala**	1.191	**Escadaria**	884	**Duque de Caxias**	2.986
Porto Novo	1.782	**Vias Principais**	1.170	**Escadaria**	912	**São João de Meriti**	2.325
Penha	1.659	**Pontos Estratégicos**	1.141	**Morro Rato**	828	**Belford Roxo**	2.397
Caonze	1.621	**Bocas de Fumo**	1.007	**Vila norma**	817	**Niterói**	1.848

A Tabela IV.7 apresenta as estatísticas descritivas do *corpus* após a rotulagem manual. Observa-se que a média de entidades por relato é 3,84, com uma mediana de 3, indicando uma distribuição assimétrica e a presença de alguns relatos com um número elevado de entidades. Além disso, 1.054 relatos não apresentam nenhuma entidade identificável, o que evidencia variabilidade significativa no conteúdo textual.

Tabela IV.7: Resumo estatístico da ocorrência de entidades nomeadas por relato no *corpus* DD_{small}. São apresentados a média, a mediana, o desvio padrão e a quantidade de relatos que não contêm entidades.

Descrição	Valores
Média de entidades por relato	3,84
Mediana de entidades por relato	3,0
Desvio padrão de entidades por relato	4,98
Quantidade de relatos sem entidades	1.054

IV.3 Etapas da metodologia

A Figura IV.2 constitui o elemento central da metodologia proposta nesta pesquisa, sintetizando o encadeamento das etapas, os fluxos de dados, os modelos empregados e as interações entre os diferentes *corpora*. A partir desse fluxo metodológico, organiza-se uma abordagem composta por duas etapas complementares, integrando estratégias de aprendizado supervisionado e semissupervisionado para o reconhecimento de entidades nomeadas em relatos criminais informais.

Na primeira etapa, utiliza-se o *corpus* rotulado DD_{small} como base supervisionada para a preparação dos dados, a anotação das entidades, a divisão entre conjuntos de treino e teste e o treinamento inicial do modelo. Essa fase tem como objetivo construir um modelo de referência especializado no domínio dos relatos criminais, capaz de identificar entidades dos tipos **Pessoa**, **Localização** e **Organização**.

Na segunda etapa, o modelo obtido é aplicado ao *corpus* DD_{large}, significativamente maior e sem anotações manuais, com o propósito de gerar pseudo-rótulos por meio de uma estratégia de auto-treinamento iterativo. Esse processo permite incorporar novos exemplos ao treinamento, ampliando a cobertura do modelo e reduzindo os efeitos da escassez de dados rotulados.

De forma transversal a todo o fluxo apresentado na Figura IV.2, os metadados estruturados associados aos relatos são empregados como fonte complementar de contexto, auxiliando tanto o REN quanto a validação e o refinamento das predições geradas automaticamente. Assim, a metodologia integra o ajuste fino do modelo GLiNER sobre o DD_{small}, a utilização de pseudo-rotulagem sobre o DD_{large} e o aproveitamento de informações estruturadas, constituindo o eixo

central desta pesquisa.

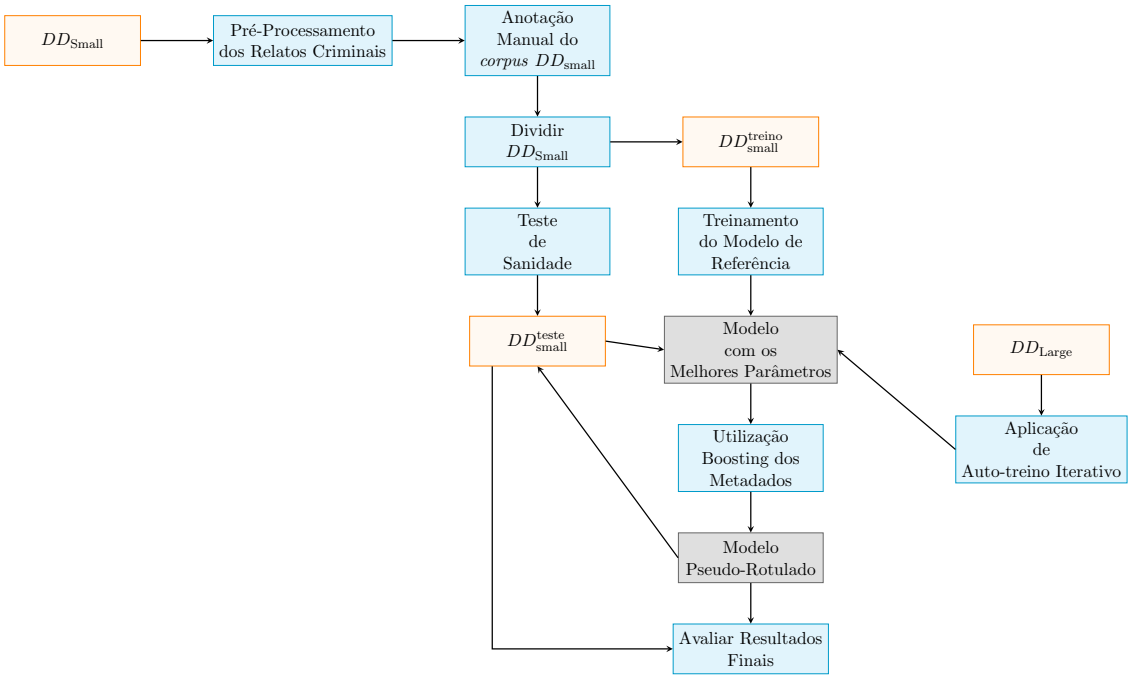


Figura IV.2: Fluxograma da metodologia proposta para reconhecimento de entidades nomeadas. Os blocos em laranja representam os *corpus* utilizados, os modelos estão indicados em cinza e os processos estão destacados em azul.

IV.3.1 Pré-Processamento dos Relatos Criminais

Os relatos criminais que compõem os *corpora* desta pesquisa apresentam predominantemente uma linguagem informal, além de diversos erros ortográficos e gramaticais. No entanto, optou-se por não realizar nenhuma correção ou padronização dessas características durante a etapa de pré-processamento. Essa decisão fundamenta-se na preservação da naturalidade dos textos, uma vez que tais particularidades refletem a forma de expressão dos denunciantes. Dessa maneira, busca-se garantir que as entidades extraídas sejam compatíveis com aquelas que possivelmente serão encontradas em *corpora* futuros, mantendo a aderência ao domínio real de aplicação.

Para fins de análise e utilização nos modelos de REN, os relatos criminais são submetidos a uma etapa de pré-processamento textual. Inicialmente, realiza a tokenização dos textos, processo que consiste na segmentação do conteúdo em unidades léxicas menores, denominadas *tokens*, que constituem a unidade básica de entrada dos modelos de linguagem. Adicionalmente, foram removidos caracteres especiais, símbolos não informativos e elementos textuais inconsistentes que poderiam comprometer a correta geração dos *tokens*, reduzindo ruídos no conjunto de dados.

Essa etapa é fundamental para padronizar a representação textual dos relatos, especialmente em um contexto caracterizado por linguagem informal, abreviações, erros ortográficos e variações de escrita. Dessa forma, busca-se preservar as informações semanticamente relevantes, ao mesmo

tempo em que se minimizam inconsistências que possam afetar o processo de treinamento e inferência dos modelos de PLN.

IV.3.2 Anotação Manual do *corpus* DD_{small}

Os *tokens* identificados como pertencentes a entidades nomeadas foram rotulados segundo as categorias previamente definidas, como pessoas, locais e organizações. Por sua vez, os *tokens* que não compõem nenhuma dessas categorias foram classificados como objeto (O), representando elementos contextuais ou complementares que, embora não sejam o foco principal da extração, contribuem para a compreensão semântica do relato.

Para a anotação de entidades no DD_{small}, adotou-se o modelo de marcação BIO, que permite representar a sequência das entidades de maneira estruturada e padronizada. Esse esquema de anotação distingue o início de uma entidade por meio do rótulo B, as palavras subsequentes pertencentes à mesma entidade pelo rótulo I, e os *tokens* que não fazem parte de nenhuma entidade nomeada pelo rótulo O. A Tabela IV.8 apresenta exemplos ilustrativos dessa marcação:

Tabela IV.8: Exemplos de marcação no formato BIO

Expressão	Marcação BIO
Rio de Janeiro	B-LOCAL, I-LOCAL, I-LOCAL
Polícia Militar	B-ORGANIZAÇÃO, I-ORGANIZAÇÃO
Marcos Paulo	B-PESSOA, I-PESSOA

A anotação das entidades foi realizada manualmente no DD_{small}, por meio da leitura individual dos relatos e da identificação das entidades mencionadas, seguindo as categorias previamente definidas. Essa etapa teve como finalidade produzir um conjunto de dados anotados de forma consistente, destinado ao treinamento dos modelos, os quais foram posteriormente avaliados no conjunto de teste, utilizando o mesmo esquema de marcação. Esse procedimento possibilitou mensurar a qualidade das abordagens testadas.

A validação manual foi conduzida por meio da leitura individual dos relatos e da verificação criteriosa das entidades nomeadas identificadas no *corpus* DD_{small}. Esse procedimento assegurou a qualidade das anotações utilizadas no treinamento quanto à avaliação dos modelos. A Tabela IV.9 sintetiza os principais resultados dessa etapa.

A Tabela IV.10 apresenta um resumo estatístico da distribuição das entidades nomeadas identificadas no corpus DD_{small} após o processo de anotação manual. Essa etapa foi conduzida com a colaboração dos pesquisadores colaboradores **Gabriel Rodrigo Bezerra Araujo** e **Natan Balthazar de Andrade**, que atuaram como anotadores humanos, sendo posteriormente revisada e validada pelo autor desta dissertação. Os valores apresentados correspondem ao número total de

Tabela IV.9: Distribuição das entidades identificadas no *corpus* utilizado na validação manual, apresentando o total de ocorrências e a quantidade de entidades de Pessoa, Localização e Organização verificadas após a revisão humana.

Descrição	Valor
Total de entidades	20.101
Entidades de Pessoa	4.020
Entidades de Localização	12.301
Entidades de Organização	3.780

entidades anotadas em cada categoria, constituindo a base supervisionada utilizada no treinamento e na avaliação dos modelos desenvolvidos.

Observa-se que a categoria *Localização* concentra a maior parcela das entidades anotadas, correspondendo a 61,2% do total (12.301 entidades). Em seguida, encontra-se a categoria *Pessoa*, com 20,0% (4.020 entidades), enquanto a categoria *Organização* representa 18,8% (3.780 entidades). A predominância de entidades relacionadas a locais e indivíduos é compatível com a natureza dos relatos criminais analisados, nos quais referências espaciais e identificação de pessoas frequentemente desempenham papel central na descrição das ocorrências.

Posteriormente, foi realizada uma segunda etapa de anotação com o objetivo de avaliar a consistência das marcações em diferentes contextos. Essa atividade contou com a participação dos pesquisadores colaboradores **João Guilherme de Oliveira Ribeiro Kongevold** e **João Pedro Jendiroba Paixão Corrêa**, que utilizaram uma plataforma especializada para anotação de entidades nomeadas. A ferramenta foi configurada de acordo com o esquema de entidades adotado neste trabalho, permitindo a marcação das categorias *Pessoa*, *Localização* e *Organização*. Após a conclusão dessa etapa, os resultados obtidos foram analisados e validados pelo autor da dissertação, contribuindo para a verificação da qualidade e da consistência das anotações realizadas.

Tabela IV.10: Resumo estatístico da distribuição das entidades

Entidades	Quantidade	Distribuição
Localização	12.301	61,2%
Organização	3.780	18,8%
Pessoa	4.020	20%

A Tabela IV.11 apresenta estatísticas descritivas detalhadas do DD_{small}, segmentadas pelas classes Localização, Pessoa e Organização. Além da quantidade total já apresentada, são exibidos o número de formas distintas registradas e os comprimentos médios das entidades. Observa-se que a classe Localização possui o maior volume de entidades (12.301), com 4.133 variações distintas

e comprimento médio de 2,22. A classe Pessoa apresenta 4.020 ocorrências, com 1.843 variações de entidades distintos e menor extensão média (1,50). Já a classe Organização, embora menos frequente (3.780 ocorrências), possui 942 variações distintas e um comprimento médio de 1,50, possivelmente devido à presença de siglas ou nomes abreviados. Esses dados são úteis para compreender o comportamento linguístico do *corpus* e para o direcionamento do ajuste fino do modelo.

Tabela IV.11: Resumo estatístico da anotação de entidades nomeadas no DD_{small}, incluindo métricas sobre a quantidade de entidades por relato e a frequência de relatos sem entidades.

Descrição	Localização	Pessoa	Organização
Quantidade de entidades	12.301	4.020	3.780
Quantidade de entidades distintas	4.133	1.843	942
Comprimento médio das entidades	2,22	1,50	1,51

IV.3.3 Divisão do *corpus* DD_{small}

Para a condução dos treinamentos e testes dos modelos, o conjunto DD_{small} foi dividido em duas partes. Essa divisão possibilitou a validação das diferentes abordagens e serviu de base para comparações entre configurações distintas, assegurando uma avaliação mais robusta do desempenho dos modelos.

A divisão adotada reservou 20% dos relatos criminais (totalizando 1.004 relatos) para a etapa de teste, enquanto os 80% restantes (equivalentes a 4.226 relatos) compuseram o conjunto de treinamento. A Tabela IV.12 apresenta a distribuição das entidades nomeadas em cada subconjunto, permitindo uma visão detalhada da composição do DD_{small} após a separação dos dados.

Tabela IV.12: Frequência de Entidades nos Conjuntos de Treino e Teste

Entidades	Treino	Teste
Localização	9.862	2.439
Organização	2.993	787
Pessoa	3.171	849

Para aprimorar a utilização do modelo GLiNER (descrito na Seção II.2.3) e tornar os testes mais rigorosos, foi adotado um teste de sanidade. Essa abordagem tem como objetivo filtrar o conjunto de teste para garantir que ele contenha apenas entidades que não aparecem no conjunto de treinamento. A aplicação desse teste tornou-se necessária após a identificação de uma grande interseção entre as entidades encontradas nos dados de treinamento e aquelas presentes na base de teste original.

Assim, o DD_{small} foi reorganizado de modo que apenas rótulos inéditos em relação ao treinamento fossem considerados na etapa de avaliação. Essa estratégia possibilita verificar o desempenho do modelo diante de informações realmente novas, simulando cenários de aplicação em domínios diferentes.

Após a aplicação do teste de sanidade, a base de teste passou a conter um total de 980 entidades, distribuídas da seguinte forma: 128 entidades da classe Organização, 570 da classe Localização e 282 da classe Pessoa.

IV.3.4 Treinamento do Modelo de Referência

A Figura IV.3 apresenta o procedimento adotado para o treinamento do modelo de referência, que consiste no ajuste fino de um modelo pré-treinado ao domínio dos relatos do Disque Denúncia, combinado com validação cruzada aninhada e otimização do limiar de predição. O processo inicia-se pela divisão do DD_{small}, conforme destacado em IV.3.3,em dois subconjuntos: um destinado ao treinamento e validação dos modelos e outro reservado exclusivamente para a avaliação final do desempenho (conjunto de teste).

O subconjunto de treinamento é então submetido a uma validação cruzada aninhada com $k = 5$. Em cada iteração da validação externa, uma partição é reservada para validação, enquanto as demais são utilizadas para treinamento. Sobre esse conjunto de treinamento, realiza-se uma validação interna, na qual diferentes configurações do modelo são avaliadas com base na métrica *F1* ao nível de entidade, conforme descrito na Seção II.2.3.

Durante a validação interna, também é realizada a otimização do limiar de predição (*threshold*), que define o valor mínimo de confiança para que uma entidade seja considerada válida. Diferentes valores de limiar são testados e comparados com base na métrica *F1*, sendo selecionado aquele que apresenta melhor desempenho.

Ao final da validação cruzada, seleciona-se o modelo associado à melhor configuração e ao melhor desempenho observado nas iterações externas. Esse modelo de referência, denotado por M_0, é então avaliado no subconjunto $DD_{\text{small}}^{\text{teste}}$, permitindo estimar sua capacidade de generalização.

IV.3.5 Aplicação de Auto-treino Iterativo

Após a obtenção do modelo de referência M_0, descrito na Seção IV.3.4, inicia-se a etapa de aprendizado semissupervisionado, cujo objetivo é explorar o grande volume de dados não rotulados disponíveis no *corpus* DD_{large}. Essa etapa busca mitigar a limitação imposta pelo tamanho reduzido do conjunto anotado DD_{small}, ampliando o conjunto de treinamento por meio de pseudo-rotulagem. Essa estratégia baseia-se no princípio de que predições de alta confiança pode ser utilizadas como aproximações dos rótulos verdadeiros, permitindo expandir o conjunto de treinamento sem anotação

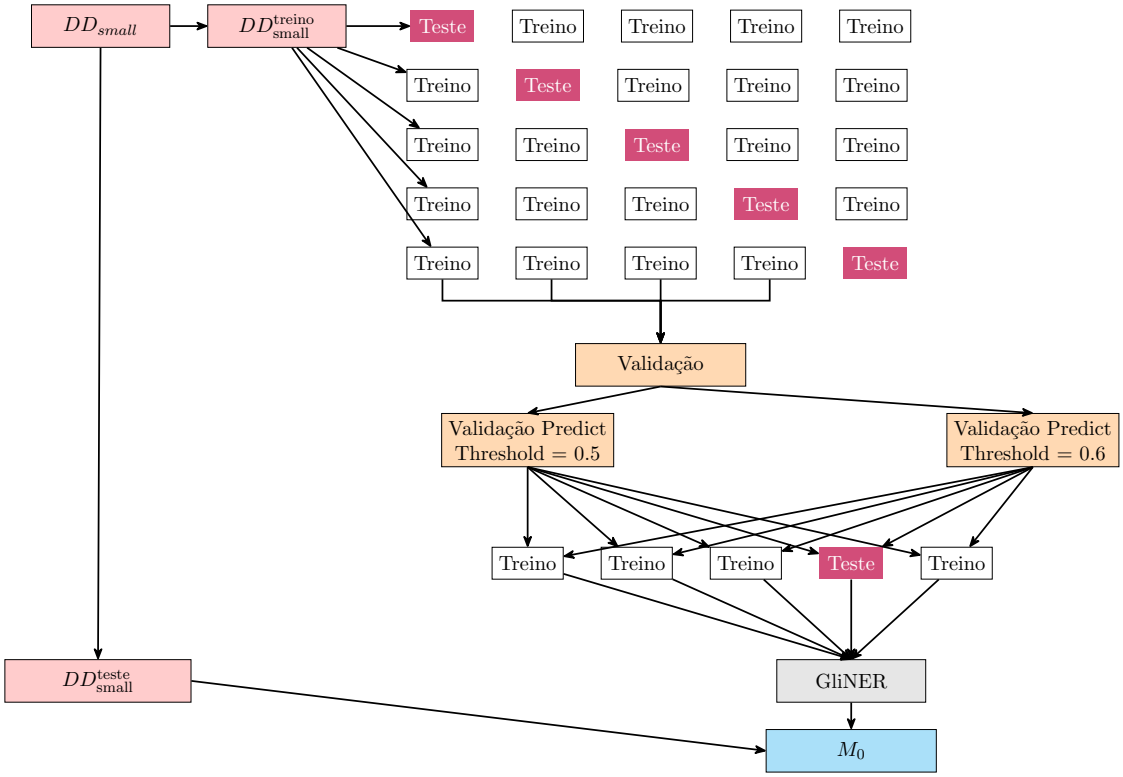


Figura IV.3: Fluxograma do processo de validação cruzada aninhada sobre o *corpus* DD_{small}. Inicialmente, o *corpus*, representado nas cores vermelhas, é dividido em subconjuntos de treino e teste. Em seguida, o subconjunto de treino, na cor branca, é utilizado em uma validação cruzada externa com cinco dobras, onde em cada iteração, um dos *folds* é reservado para teste identificado na cor roxa e as demais para treinos. Ao final da validação externa, os dados de treino são novamente particionados para realizar uma validação interna verificando novamente os *folds* presente nas cores laranja, na qual diferentes limiares de confiança são avaliados. Com base nessa avaliação, o melhor modelo é treinado utilizando os dados de treino correspondentes e avaliado com os dados de teste e então selecionado. Finalmente, esse modelo M_0 é testado no subconjunto de teste separado.

manual adicional.

Cada entidade identificada é avaliada por meio do valor de confiança atribuído pelo modelo à respectiva predição. Para aumentar a precisão da análise, esse valor é comparado a um limiar (*threshold*) previamente definido. Apenas as predições cujos valores de confiança superam esse limiar são aceitas, para reduzir falsos positivos. A confiança γ_e é calculada segundo a Equação IV.1, considerando o maior score entre todas as classes possíveis.

$$\gamma_e = \max_i \Pr(y = i \mid \mathbf{x}_e) \tag{IV.1}$$

Sendo:

- γ_e é a confiança associada à predição da entidade e;
- $\mathbf{x}_e$ – vetor de entrada correspondente à entidade e;
- $y = i$ – classe prevista pelo modelo para a entidade;

- $\Pr(y = i \mid \mathbf{x}_e)$ – probabilidade de a entidade pertencer à classe i;
- $\max_i$ – maior probabilidade entre todas as classes possíveis.

O Algoritmo 1 apresenta o processo de auto-treinamento iterativo adotado neste trabalho. Inicialmente, um modelo M_0 é ajustado por meio de ajuste fino sobre o *corpus* DD_{small}, servindo como ponto de partida para a pseudo-rotulagem. Em cada iteração, o modelo é aplicado ao $DD_{\text{large não_rot}}$ (inicialmente equivalente ao *corpus* DD_{large}) para gerar predições $\hat{y}$ e seus respectivos valores de confiança γ_e conforme a Equação IV.1. Quando uma entidade reconhecida e metadados estruturados do relato são semelhantes, a confiança associada é ajustada para $\tilde{\gamma}_e$, conforme a Equação IV.3. Apenas as predições que satisfazem o critério $\tilde{\gamma}_e \geq \tau$ (Limiar de Confiança da Entidade) são incorporadas ao novo conjunto de treinamento.

Nesta etapa, o modelo é refinado exclusivamente com os exemplos pseudo-rotulados cuja confiança ajustada supera o limiar τ, sem reincorporação do conjunto original DD_{small}. Esse processo é repetido por T iterações e em cada uma dessas é adotada um limiar diferente para cada relato criminal, permitindo que o modelo explore progressivamente exemplos informativos não rotulados e aprimore sua capacidade de generalização com base em predições consideradas confiáveis. Após cada iteração, os exemplos utilizados no treinamento são removidos de $DD_{\text{large não_rot}}$. Essa estratégia evita a reutilização de exemplos já explorados, reduzindo o risco de reforço de erros e promovendo maior diversidade no conjunto de treinamento ao longo das iterações.

A Figura IV.4 ilustra o fluxo metodológico empregado na etapa de auto-treinamento iterativo. Inicialmente, utiliza-se o DD_{large}, composto por um grande volume de relatos criminais não rotulados. Esse conjunto é automaticamente anotado pelo modelo M_0, previamente selecionado, na etapa de treinamento descrita na Seção IV.3.4.

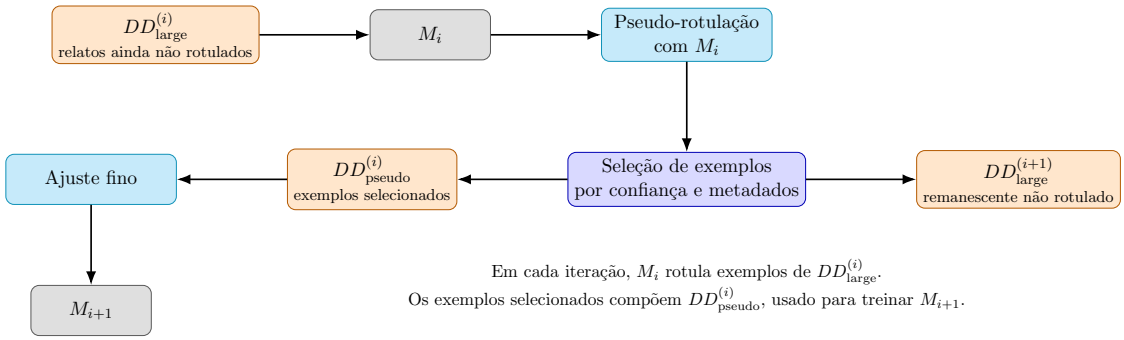


Figura IV.4: Fluxo do processo de auto-treinamento iterativo. Em cada iteração, o modelo corrente M_i é aplicado ao conjunto ainda não rotulado $DD^{(i)}_{\text{large}}$ para gerar pseudo-rótulos. Após a seleção dos exemplos mais confiáveis, apoiada por escores de confiança e metadados estruturados, obtém-se o conjunto $DD^{(i)}_{\text{pseudo}}$, utilizado no ajuste fino que produz o modelo seguinte, M_{i+1}. Os exemplos já selecionados são removidos do conjunto não rotulado, formando $DD^{(i+1)}_{\text{large}}$.

A métrica de *loss* adotada é a entropia cruzada categórica (*categorical cross-entropy loss*),

Algorithm 1 Auto-treinamento iterativo com pseudo-rotulagem, calibração opcional, ajuste por metadados e atualização do limiar

Require: Modelo inicial M_0, *corpus* não rotulado DD_{large}, número de iterações T, limiar de confiança τ, fator de ajuste α, método de calibração $\mathcal{C} \in \{\text{RAW}, \text{TS}, \text{IR}\}$
Ensure: Modelo final M_T

$DD_{\text{large}}^{(0)} \leftarrow DD_{\text{large}}$
for $i = 0$ **até** $T - 1$ **do**
  $DD_{\text{pseudo}}^{(i)} \leftarrow \emptyset$
  Definir τ_i ▷ Limiar de confiança adotado pelo pesquisador para a iteração i
  for all relato criminal $r \in DD_{\text{large}}^{(i)}$ **do**
    $\hat{Y}_r \leftarrow M_i(r)$ ▷ Entidades preditas para o relato criminal r
    for all entidade $e \in \hat{Y}_r$ **do**
      $\gamma_e \leftarrow \text{confianca}_e$ ▷ Confiança original conforme Eq. IV.1
      if $\mathcal{C} = \text{RAW}$ **then**
        $\gamma_e^{\text{cal}} \leftarrow \gamma_e$
      else if $\mathcal{C} = \text{TS}$ **then**
        $\gamma_e^{\text{cal}} \leftarrow \gamma_e^{\text{TS}}$ ▷ Confiança calibrada por *Temperature Scaling*
      else if $\mathcal{C} = \text{IR}$ **then**
        $\gamma_e^{\text{cal}} \leftarrow \gamma_e^{\text{IR}}$ ▷ Confiança calibrada por *Isotonic Regression*
      end if
      if $e \in \text{Metadados}(r)$ **then** ▷ Comparação entre a entidade e os metadados descritos na Seção IV.2
        $\tilde{\gamma}_e \leftarrow \min(\alpha \cdot \gamma_e^{\text{cal}}, 1)$ ▷ Aplicação do fator de ajuste conforme Eq. IV.3
      else
        $\tilde{\gamma}_e \leftarrow \gamma_e^{\text{cal}}$
      end if
    end for
    if $|\hat{Y}_r| > 0$ **then**
      $\gamma_{\text{relato}} \leftarrow \frac{1}{|\hat{Y}_r|} \sum_{e \in \hat{Y}_r} \tilde{\gamma}_e$ ▷ Confiança do relato criminal conforme Eq. IV.9
    end if
    if $\gamma_{\text{relato}} \geq \tau_i$ **then**
      $DD_{\text{pseudo}}^{(i)} \leftarrow DD_{\text{pseudo}}^{(i)} \cup \{(r, \hat{Y}_r)\}$
    end if
  end for
  $M_{i+1} \leftarrow DD_{\text{pseudo}}^{(i)}$ ▷ Inclusão dos pseudo-rótulos selecionados no treinamento através de ajuste fino
  $DD_{\text{large}}^{(i+1)} \leftarrow DD_{\text{large}}^{(i)} - DD_{\text{pseudo}}^{(i)}$ ▷ Remove os relatos incorporados ao treinamento
end for
return M_T

utilizada para mensurar a diferença entre a classe verdadeira de cada exemplo e a distribuição de probabilidades prevista pelo modelo. No contexto deste trabalho, considera-se como exemplo cada instância avaliada pelo modelo durante o treinamento, isto é, uma entidade candidata, um trecho textual ou uma unidade de entrada processada pelo modelo, conforme a formulação adotada na tarefa de REN.

A função de perda é apresentada na Equação IV.2:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log(\hat{y}_{i,c}). \tag{IV.2}$$

Sendo:

- i – índice do exemplo avaliado, em que cada exemplo corresponde a uma instância de entrada utilizada no treinamento;
- N – número total de exemplos considerados no cálculo da perda;
- C – número total de classes possíveis;
- $y_{i,c}$ – rótulo verdadeiro do exemplo i para a classe c, representado em formato *one-hot encoding*. Assim, $y_{i,c} = 1$ quando a classe c é a classe correta do exemplo i, e $y_{i,c} = 0$ caso contrário;
- $\hat{y}_{i,c}$ – probabilidade prevista pelo modelo de que o exemplo i pertença à classe c.

Na prática, como o rótulo verdadeiro é representado em formato *one-hot*, apenas a classe correta contribui para o valor da perda. Por exemplo, se um exemplo pertence à classe `Pessoa`, então o vetor verdadeiro pode ser representado como $(1, 0, 0)$ para as classes `Pessoa`, `Localização` e `Organização`. Nesse caso, a perda considera apenas o logaritmo da probabilidade prevista para a classe `Pessoa`. Portanto, quanto maior a probabilidade atribuída pelo modelo à classe correta, menor será o valor da perda.

A análise combinada da função de perda e da métrica de confiança possibilita uma avaliação mais completa do desempenho do modelo. A função de perda fornece uma medida quantitativa da discrepância entre as predições geradas e os rótulos verdadeiros, enquanto a métrica de confiança expressa o grau de certeza associado às decisões do modelo. Em conjunto, essas medidas permitem monitorar a evolução do aprendizado, avaliar a capacidade de generalização e apoiar a seleção de exemplos pseudo-rotulados de maior confiabilidade. Consequentemente, fornecem subsídios para uma análise mais consistente do comportamento do modelo e para sua utilização em diferentes *corpora* de relatos criminais.

Aplicação de Fator de Ajuste

Durante o processo de pseudo-rotulagem, as entidades reconhecidas foram confrontadas com os metadados estruturados associados a cada relato. Para cada entidade predita, verificou-se a ocorrência de correspondência com os campos *bairro*, *logradouro*, *ponto de referência* e *cidade*. A presença dessa correspondência foi interpretada como uma evidência contextual adicional em favor da predição, indicando consistência entre a informação extraída do texto livre e os dados estruturados disponíveis. Nessas situações, aplicou-se um incremento de 20% ao fator de ajuste γ_e, utilizado no cálculo da confiança da entidade. A definição desse percentual ocorreu de forma heurística, inspirada no princípio de Pareto, de modo a atribuir maior peso às entidades corroboradas pelos metadados sem introduzir distorções excessivas no processo de calibração.

De maneira mais formal, para cada entidade e reconhecida automaticamente em DD_{large}, calcula-se a confiança γ_e conforme a Equação IV.1. Caso haja confirmação por metadados estruturados, aplica-se o fator $\alpha = 1{,}2$, garantindo que a confiança ajustada $\tilde{\gamma}_e$ permaneça normalizada no intervalo $[0, 1]$.

Essa estratégia visa refletir a maior confiabilidade das entidades validadas por dados externos estruturados, contribuindo para o aumento da assertividade do modelo durante o refinamento.

$$\tilde{\gamma_e} = \min\left(\gamma_e \times \alpha,\ 1\right) \tag{IV.3}$$

Nesta dissertação, adota-se $\alpha = 1{,}2$ como fator de ajuste da confiança em entidades confirmadas por metadados, conforme indicado na Equação IV.3. Esse valor foi escolhido de forma heurística, com base nos experimentos utilizados na Seção V.7.

Esse valor de incremento de 20% é adotado por representar um reforço moderado às predições corroboradas por informações estruturadas, as quais fornecem evidências contextuais adicionais e independentes do texto livre. Tal estratégia está alinhada à proposta metodológica de integração entre dados textuais e metadados, visando aumentar a confiabilidade das entidades reconhecidas.

Do ponto de vista empírico, observa-se que fatores de ajuste inferiores produzem impacto pouco significativo na discriminação entre predições confiáveis e não confiáveis. Por outro lado, valores mais elevados resulta em uma amplificação excessiva das pontuações de confiança, comprometendo a calibração do modelo e elevando a probabilidade de incorporação de falsos positivos durante o processo de pseudo-rotulagem.

Dessa forma, o valor adotado para o fator α configura um compromisso entre a valorização das evidências provenientes dos metadados e a preservação das propriedades probabilísticas das estimativas de confiança, contribuindo para a estabilidade e a robustez do processo de auto-treinamento iterativo.

Entretanto, a utilização direta das estimativas de confiança pode ser problemática, uma vez que modelos de linguagem frequentemente apresentam comportamento sobreconfiantes,atribuindo escores elevados a predições incorretas. Nesse contexto, a calibração (Seção II.3.3) torna-se essencial para garantir que os valores de confiança reflitam os valores de confiança às probabilidades reais de acerto do modelo, garantindo uma interpretação mais fiel da incerteza associada às predições. Esse fenômeno de má calibração em redes neurais profundas é amplamente discutido na literatura, conforme evidenciado por Guo et al. [2017], que demonstram a necessidade de técnicas específicas para corrigir tais desvios.

Calibração dos Modelos

Para verificar e validar a confiabilidade das estimativas de confiança produzidas pelo modelo treinado, realiza-se uma análise de calibração com o objetivo de identificar possíveis casos de sobre-confiança, caracterizados por baixa variabilidade das pontuações $\gamma_e \in [0, 1]$ e pouca separabilidade entre exemplos. Essa situação pode comprometer as etapas subsequentes do processo autoiterativo. Nesse contexto, analisaram-se a distribuição de $\{\gamma_e\}_{e=1}^{n}$, a capacidade discriminativa entre classes e métricas específicas de calibração, de modo a verificar se as probabilidades previstas refletem adequadamente a incerteza do modelo. Esse procedimento é fundamental no contexto desta metodologia, que depende diretamente da qualidade das estimativas de confiança para a seleção de pseudo-rótulos.

Para construir o diagnóstico de calibração, considera-se diretamente o conjunto de pares formados pelas estimativas de confiança $\gamma_e \in [0, 1]$ e pelas predições correspondentes. A partir da Equação IV.1, em que γ_e representa a confiança *top-1* do exemplo e, define-se a predição de acordo com a Equação IV.4 como:

Seguindo a notação da Equação IV.1, em que γ_e representa a confiança *top-1* atribuída à entidade e, define-se a classe predita $\hat{y}_e$ como aquela associada à maior probabilidade estimada pelo modelo:

$$\hat{y}_e = \arg\max_i \Pr(y = i \mid \mathbf{x}_e). \tag{IV.4}$$

Nessa formulação, $\mathbf{x}_e$ representa a entrada associada à entidade e, i percorre o conjunto de classes possíveis, e $\hat{y}_e$ corresponde ao rótulo escolhido pelo modelo. Para avaliar a calibração, compara-se a confiança prevista γ_e com o resultado observado da predição, representado pelo indicador $\mathbf{1}[\hat{y}_e = y_e]$. Esse indicador assume valor 1 quando a classe predita $\hat{y}_e$ coincide com o rótulo verdadeiro y_e, indicando uma predição correta, e valor 0 caso contrário.

Assim, a avaliação da calibração conduzida pela comparação entre as probabilidades previstas e os acertos observados. Um modelo bem calibrado deve apresentar alinhamento entre essas quan-

tidades, de modo que valores mais elevados de confiança estejam associados a maiores proporções de predições corretas.

O diagrama de confiabilidade (*reliability diagram*) utiliza-se como ferramenta visual para inspecionar a calibração das estimativas produzidas pelo modelo. Nesse gráfico, o eixo horizontal representa os valores de confiança previstos (γ_e), enquanto o eixo vertical representa a frequência observada de acertos, obtida a partir do indicador $\mathbf{1}[\hat{y}_e = y_e]$. A condição ideal de calibração é representada pela reta diagonal $y = x$, na qual a confiança prevista coincide com a probabilidade empírica de acerto. Assim, por exemplo, predições realizadas com confiança de $0,8$ deveriam estar corretas em aproximadamente 80% dos casos. Quanto mais próximos os pontos estiverem dessa diagonal, melhor calibrado estará o modelo.

Adicionalmente, o erro de calibração esperado (*Expected Calibration Error* (ECE)) é empregado como métrica agregada para quantificar a discrepância entre as confianças previstas e as proporções observadas de acerto ao longo do conjunto de exemplos.

As métricas de calibração são obtidas por meio dos métodos de pós-processamento *Temperature Scaling* (TS) (*Temperature Scaling*) e *Isotonic Regression* (IR) (*Isotonic Regression*), aplicados às estimativas de confiança não calibradas definidas na Equação IV.1. No caso do TS, busca-se estimar um parâmetro escalar positivo $T > 0$, denominado temperatura, utilizando um conjunto de validação. Esse parâmetro é obtido pela minimização da perda de log-verossimilhança negativa (NLL), conforme apresentado na Equação IV.5.

$$T^{\star} = \arg\min_{T>0} \sum_{e=1}^{n_{\text{val}}} -\log\left(\text{softmax}\left(\frac{\mathbf{z} * e}{T}\right) * y_e\right), \tag{IV.5}$$

em que $T^{\star}$ representa o valor ótimo da temperatura, n_{val} corresponde ao número de exemplos do conjunto de validação, $\mathbf{z}_e$ denota o vetor de *logits* associado ao exemplo e e y_e representa sua classe verdadeira. Os *logits* consistem nos escores brutos produzidos pelo modelo antes da aplicação da função *softmax*, sendo posteriormente convertidos em probabilidades.

Após a estimação da temperatura ótima, as probabilidades calibradas são calculadas por meio da reescalagem dos *logits*, conforme descrito na Equação IV.6.

$$p^{\text{TS}}(y = i \mid \mathbf{x}_e) = \text{softmax}\left(\frac{\mathbf{z}_e}{T^{\star}}\right)_i, \qquad \gamma_e^{\text{TS}} = \max_i p^{\text{TS}}(y = i \mid \mathbf{x}_e) \tag{IV.6}$$

Nessa equação, $p^{\text{TS}}(y = i \mid \mathbf{x}_e)$ representa a probabilidade calibrada da classe i para o exemplo $\mathbf{x}_e$, enquanto γ_e^{TS} corresponde à confiança calibrada da predição, obtida pela maior probabilidade entre todas as classes possíveis. Valores de temperatura superiores a um ($T > 1$) tendem a suavizar a distribuição de probabilidades e reduzir a sobreconfiança do modelo, ao passo que valores no intervalo $0 < T < 1$ produzem distribuições mais concentradas.

Na IR, aprende-se uma função isotônica não decrescente $g : [0, 1]! \rightarrow![0, 1]$ (via *pool-adjacent-violators*) que mapeia as confianças brutas para confianças calibradas:

$$g^{\star} = \arg \min_{g,\text{isotônica}} ; \sum_{e=1}^{n_{\text{val}}} \big(g(\gamma_e) - \mathbf{1}[\hat{y}_e = y_e]\big)^2, \qquad \gamma_e^{\text{IR}} = g^{\star}(\gamma_e). \tag{IV.7}$$

Após a calibração (TS ou IR), aplicamos o fator de ajuste de confiança especificado $\alpha \in (0, \infty)$ para obter as pontuações finais:

$$\tilde{\gamma}_e = \min \Big\{ \alpha \cdot \gamma_e^{\text{cal}}, 1 \Big\} \tag{IV.8}$$

onde $\gamma_e^{\text{cal}} \in \gamma_e^{\text{TS}}, \gamma_e^{\text{IR}}$. A operação de truncamento assegura que a confiança ajustada permaneça no intervalo $[0, 1]$. As curvas de confiabilidade e o ECE foram calculadas sobre o conjunto $\tilde{\gamma}_e$.

Seleção de Relatos para Pseudo-rotulagem

Durante o processo de auto-treinamento iterativo, é necessário selecionar quais exemplos pseudo-rotulados serão incorporados ao conjunto de treinamento em cada iteração. Para isso, propõe-se nesta dissertação uma métrica de confiança em nível de relato, denominada *confiança do relato*, desenvolvida pelo autor com o objetivo de agregar as informações de confiança das entidades reconhecidas em uma única medida representativa da qualidade global das predições produzidas para cada relato criminal.

Seja um relato contendo n entidades identificadas, cada uma com confiança γ_{e_i}. A confiança global do relato é definida como a média das confianças dessas entidades:

$$\gamma_{\text{relato}} = \frac{1}{n} \sum_{i=1}^{n} \gamma_{e_i} \tag{IV.9}$$

Essa formulação foi proposta especificamente para esta pesquisa com a finalidade de avaliar a confiabilidade de um relato como um todo, em contraste com a análise individual de cada entidade reconhecida. A hipótese adotada é que relatos cujas entidades apresentam, em média, maiores níveis de confiança tendem a conter predições mais consistentes e, consequentemente, oferecem menor risco de propagação de erros durante o processo de pseudo-rotulagem.

Em cada iteração do auto-treinamento, apenas os relatos cuja confiança global excede um limiar pré-definido são selecionados para compor o conjunto pseudo-rotulado DD_{pseudo}. Esse conjunto é então utilizado para realizar uma nova etapa de ajuste fino do modelo, permitindo sua atualização com base em exemplos adicionais provenientes de dados originalmente não rotulados.

A utilização de um ponto de corte sobre a confiança do relato busca reduzir a incorporação de exemplos potencialmente incorretos ao treinamento, priorizando relatos com maior grau de

confiabilidade. Dessa forma, torna-se possível expandir progressivamente o conjunto de treinamento ao longo das iterações, mantendo maior controle sobre a qualidade dos pseudo-rótulos gerados.

IV.3.6 Avaliação

A partir da versão pseudo-rotulada do $DD_{\textbf{large}}$ realiza-se uma nova etapa de refinamento por meio de ajuste fino, avaliando diferentes níveis de confiança para determinar o limiar que oferece o melhor desempenho entre as métricas consideradas. Os modelos resultantes são comparados àqueles obtidos na validação cruzada aninhada, permitindo analisar os ganhos (ou perdas) decorrentes da inclusão de pseudo-rótulos. Por fim, os resultados são consolidados para aferir a efetividade da pseudo-rotulagem no aprimoramento do modelo.

Para avaliar o desempenho dos modelos ao longo do processo, tanto no refinamento durante a validação cruzada quanto na pseudo-rotulagem automática, foram adotadas métricas específicas que orientam a seleção dos melhores parâmetros e asseguram maior precisão na identificação das entidades nomeadas.

A seleção dos melhores modelos baseia-se na análise das métricas de *loss* e confiança. Esses critérios permitem identificar as configurações mais adequadas para o processo de pseudo-rotulagem, contribuindo para a ampliação e o aprimoramento do *corpus* $DD_{\textbf{large}}$ empregado no treinamento semissupervisionado.

Conforme descrito na Seção IV.2, o *corpus* DD_{small} é composto por relatos criminais cujas entidades nomeadas foram previamente identificadas e validadas por anotadores humanos. Em razão dessa característica, esse *corpus* serve como referência para a avaliação dos modelos desenvolvidos nesta dissertação.

A avaliação é conduzida de forma sistemática ao longo das diferentes etapas da metodologia proposta, abrangendo tanto o modelo inicial quanto os modelos posteriormente refinados por meio da estratégia de pseudo-rotulagem e da incorporação do fator de ajuste baseado nos metadados estruturados. O objetivo consiste não apenas em mensurar os ganhos globais de desempenho, mas também em analisar o impacto da calibração das estimativas de confiança e do fator de ajuste na qualidade das entidades reconhecidas.

A análise é estruturada em duas perspectivas complementares: uma avaliação quantitativa, fundamentada em métricas objetivas de desempenho ao nível de entidade, e uma avaliação qualitativa, voltada à interpretação dos resultados e à análise dos erros cometidos pelos modelos.

Avaliação Quantitativa

A avaliação quantitativa é realizada ao nível de entidade, considerando-se correta apenas a predição cuja delimitação textual e cuja classe atribuída coincidam integralmente com a anotação

de referência. Esse procedimento adota um critério rigoroso de correspondência exata, amplamente utilizado em tarefas de reconhecimento de entidades nomeadas, assegurando uma avaliação precisa da capacidade dos modelos em identificar corretamente tanto os limites quanto a categoria das entidades.

Foram adotadas as métricas de Precisão, Revocação e *F1*, amplamente utilizadas em tarefas de rotulação de sequência. A Precisão quantifica a proporção de entidades corretamente previstas entre todas as entidades identificadas pelo modelo, refletindo sua capacidade de evitar falsos positivos. A Revocação mensura a proporção de entidades reais corretamente recuperadas, indicando a capacidade do modelo de minimizar falsos negativos. O *F1*, por sua vez, corresponde à média harmônica entre Precisão e Revocação, fornecendo uma medida balanceada de desempenho.

As métricas são calculadas individualmente para cada classe (Localização, Pessoa e Organização) e também de forma agregada, permitindo observar possíveis assimetrias entre categorias. Considerando a predominância de entidades de Localização no *corpus*, essa análise estratificada foi essencial para evitar que o desempenho global mascarasse dificuldades específicas em classes menos frequentes.

Avaliação Qualitativa

A avaliação qualitativa baseia-se na inspeção manual de um subconjunto representativo de predições incorretas no conjunto de teste do DD_{small}. Os erros são organizados nas seguintes categorias:

- **Falsos Positivos (FP)**: entidades identificadas pelo modelo, mas ausentes nas anotações manuais;
- **Falsos Negativos (FN)**: entidades presentes no texto que não foram reconhecidas;
- **Erros de Classificação**: entidades corretamente delimitadas, porém atribuídas à classe incorreta;
- **Erros associados à informalidade**: variações ortográficas, abreviações e construções não padronizadas.

A análise qualitativa teve como objetivo investigar os efeitos da pseudo-rotulagem ajustada por metadados, examinando de que forma a confirmação por informações estruturadas influencia os valores de confiança atribuídos às entidades reconhecidas. Buscou-se avaliar se a incorporação desses metadados atua como um mecanismo de reforço das predições, aumentando a confiabilidade das entidades selecionadas para o processo de auto-treinamento. Adicionalmente, foram analisados

possíveis impactos decorrentes de inconsistências, ambiguidades ou incompletudes presentes nos campos estruturados, bem como sua influência na qualidade das predições geradas.

Assim, a combinação entre avaliação quantitativa e qualitativa fornece uma visão abrangente do comportamento do modelo, permitindo não apenas medir seu desempenho, mas compreender suas limitações estruturais e linguísticas no contexto real de aplicação.

Capítulo V Experimentos Computacionais

Neste capítulo, apresenta-se uma análise experimental detalhada do desempenho e da eficácia do método proposto. Cada seção discute aspectos específicos da avaliação, abrangendo desde a análise exploratória dos dados até comparações com métodos amplamente estabelecidos na literatura. A Seção V.2 descreve a análise exploratória dos dados utilizados nos experimentos. Em seguida, a Seção V.3 apresenta a configuração experimental, incluindo o ambiente computacional e as bibliotecas Python empregadas, assegurando a reprodutibilidade dos resultados. Na Seção V.4, o modelo proposto é avaliado por meio de comparações com diferentes alternativas e combinações de parâmetros testados ao longo dos experimentos. A Seção V.5 discute, de forma aprofundada, a comparação entre o método apresentado e a linha de base construída a partir de uma arquitetura LSTM, além de outros modelos avaliados, permitindo uma análise crítica dos ganhos obtidos com a abordagem adotada nesta pesquisa. A Seção V.6 descreve a aplicação das técnicas *Isotonic Regression* e *Temperature Scaling* aos *corpora* utilizados, bem como a avaliação de seu impacto, analisando se essas abordagens proporcionam ganhos de desempenho e contribuem para reduzir o comportamento de superconfiança observado nas predições. Por fim, a Seção V.7 examina o impacto da técnica de pseudo-rotulação, demonstrando como sua aplicação pode contribuir para a ampliação do dicionário de entidades e para o aprimoramento do desempenho dos modelos.

V.1 Plano Experimental

Este capítulo apresenta o plano de experimentos elaborado para validar a metodologia desenvolvida nesta dissertação. O objetivo central é avaliar, de forma sistemática, a efetividade do uso combinado do ajuste fino do modelo GLiNER, da pseudo-rotulagem e da incorporação de metadados estruturados na tarefa de reconhecimento de entidades nomeadas em relatos criminais informais provenientes do Disque Denúncia.

1. **Experimentos planejados**

   a) **Treinamento-base (baseline)**

   - Realização do ajuste fino do modelo GLiNER sobre o conjunto DD_{small}, sem utilização de metadados estruturados e sem aplicação de pseudo-rotulagem.

- Estabelecimento de uma referência inicial de desempenho para comparação com os cenários experimentais subsequentes.

b) **Impacto dos metadados**

- Inclusão de metadados estruturados (bairro, cidade, logradouro, ponto de referência) como contexto adicional fornecido ao modelo durante o treinamento.
- Avaliação do ganho relativo de desempenho em comparação com o modelo baseline.

c) **Pseudo-rotulagem iterativa**

- Aplicação do modelo ajustado sobre DD_{small} para rotular automaticamente o *corpus* DD_{large}.
- Seleção de pseudo-rótulos com base em limiares de confiança.
- Realização de ciclos iterativos de retreinamento, acompanhados da análise da evolução do desempenho ao longo das iterações.

d) **Comparação entre modelos**

- Avaliação comparativa entre o GLiNER e outros modelos de referência, incluindo BERTimbau, Spacy, LSTM e modelos de linguagem de propósito geral (GPT e Gemini), utilizando os mesmos *corpora* e o mesmo protocolo experimental.

e) **Calibração dos modelos**

- Aplicação de técnicas de calibração, como *Temperature Scaling* e *Isotonic Regression*, com o objetivo de alinhar probabilidades preditas às frequências observadas.
- Investigação do impacto dessas técnicas na confiabilidade das predições e na utilidade prática dos escores gerados.

2. **Métricas de avaliação**

- **Precisão, Revocação e F1**: métricas fundamentais para mensurar a qualidade da detecção de entidades.
- **Curvas de confiabilidade** e **Expected Calibration Error**: utilizadas para avaliar o grau de calibração das probabilidades estimadas.
- **Análises qualitativas**: inspeção manual de padrões de erros recorrentes, com ênfase em falsos positivos relacionados a apelidos e descrições informais de locais.

V.2 Análise Exploratória

V.2.1 Rotulação dos Relatos Criminais da Validação Manual

Os rótulos atribuídos às entidades observadas nos relatos criminais desempenham um papel fundamental na ampliação da capacidade do modelo em reconhecer locais mencionados no Estado do Rio de Janeiro. Muitos desses locais são amplamente conhecidos pela população, embora não estejam formalmente registrados em bases de dados estruturadas. Essa característica torna essencial a inclusão de entidades geográficas informais, contribuindo para um modelo mais sensível às particularidades linguísticas presentes nesse tipo de *corpus*. A Tabela V.1 apresenta exemplos ilustrativos de entidades de Localização rotuladas no formato BIO.

Tabela V.1: Exemplos de *tokens* anotados no formato BIO para a categoria Local, ilustrando como diferentes partes de nomes geográficos e expressões espaciais são segmentadas e rotuladas no processo de reconhecimento de entidades.

Exemplos	Marcação
Morro	B-LOCAL
Casa	I-LOCAL
Branca	I-LOCAL
Gogó	B-LOCAL
da	I-LOCAL
ema	I-LOCAL
comunidade	B-LOCAL
águia	I-LOCAL
de	I-LOCAL
ouro	I-LOCAL

Para a categoria Pessoas, exemplificada na Tabela V.2, são considerados tanto nomes próprios quanto formas informais de identificação, como apelidos e alcunhas, amplamente presentes em relatos criminais. Essa abordagem busca não apenas apoiar a atuação das forças de segurança na identificação dos indivíduos denunciados, mas também aprimorar a compreensão das dinâmicas sociais e contextuais descritas nos relatos.

Na categoria Organizações, ilustrada na Tabela V.3, foram identificadas entidades referentes a facções criminosas, unidades policiais e empresas conhecidas. Observou-se que muitos batalhões são mencionados de acordo com a região de atuação, conforme relatado pelos denunciantes. Além disso, diferentes formas de referência à Polícia Militar, incluindo abreviações, variações linguísticas e designações informais, foram devidamente incorporadas ao processo de rotulação.

Para realizar uma avaliação mais rigorosa do experimento, foi aplicado o Teste de Sanidade,

Tabela V.2: Exemplos de *tokens* anotados no formato BIO para a categoria Pessoa, incluindo nomes próprios e referências informais frequentemente encontradas em relatos criminais. A tabela demonstra como cada termo é rotulado de acordo com sua função na entidade.

Exemplos	Marcação
Milicianos	B-PESSOA
Paula	B-PESSOA
conhecida	O
como	O
Paulete	B-PESSOA
Ricardo	B-PESSOA
Vulgo	O
Dininho	B-PESSOA

Tabela V.3: Exemplos de *tokens* anotados no formato BIO para a categoria Organização, incluindo nomes de instituições, plataformas digitais e siglas presentes nos relatos criminais. A tabela ilustra como cada termo é segmentado e classificado segundo o esquema BIO.

Exemplos	Marcação
polícia	B-ORGANIZAÇÃO
YouTube	B-ORGANIZAÇÃO
Facebook	B-ORGANIZAÇÃO
Comando	B-ORGANIZAÇÃO
Vermelho	I-ORGANIZAÇÃO

que, conforme descrito na Seção IV.3.3, assegura que as entidades utilizadas no treinamento não sejam reapresentadas no conjunto de teste empregado para avaliação. Dessa forma, garante-se que o desempenho observado reflita, de fato, a capacidade de generalização do modelo, evitando qualquer sobreposição entre dados de treino e teste.

V.2.2 Rotulação dos Relatos Criminais da Pseudo-rotulação

Após o processo de ajuste fino, foram identificados novos rótulos gerados a partir das predições do modelo aplicado ao *corpus* $DD_{\textbf{small}}$. Esses rótulos correspondem às entidades observadas nos relatos criminais e ampliam o conjunto previamente anotado de forma manual.

A Tabela V.4 e apresenta exemplos ilustrativos das entidades de Localização rotuladas no formato BIO, evidenciando a diversidade e a natureza das anotações produzidas durante essa etapa.

Para a categoria Pessoas, a Tabela V.5 apresenta exemplos das entidades de pessoa rotuladas no *corpus* $DD_{\textbf{large}}$. Esses exemplos ilustram nomes e identificadores pessoais mencionados nos

Tabela V.4: Exemplos de *tokens* anotados no formato BIO para a categoria Local após o processo de ajuste fino do modelo. A tabela ilustra como nomes de ruas, avenidas e regiões são segmentados e rotulados, evidenciando melhorias na identificação de padrões geográficos nos relatos criminais.

Exemplos	Marcação
Rua	B-LOCAL
Carioca	I-LOCAL
Região	B-LOCAL
Rua	B-LOCAL
Juvenal	I-LOCAL
Valadares	I-LOCAL
Avenida	B-LOCAL
Brasil	I-LOCAL
Rua	B-LOCAL
Alagoas	I-LOCAL

relatos criminais, destacando as ocorrências reconhecidas pelo modelo durante o processo de pseudo-rotulação.

Tabela V.5: Exemplos de *tokens* anotados no formato BIO para a categoria Pessoa após o processo de ajuste fino do modelo. A tabela apresenta tanto nomes próprios quanto termos utilizados informalmente para se referir a indivíduos nos relatos criminais, refletindo a adaptação do modelo ao vocabulário característico do *corpus*.

Exemplos	Marcação
Cícero	B-PESSOA
Paula	B-PESSOA
Paulete	I-PESSOA
Ricardo	B-PESSOA
Dininho	B-PESSOA
P2	B-PESSOA
Rafael	B-PESSOA
Biel	B-PESSOA

Para os rótulos da categoria Organizações, após o processo de ajuste fino, a Tabela V.6 apresenta exemplos das entidades identificadas no *corpus* $DD_{\textbf{large}}$. Esses exemplos ilustram organizações mencionadas nos relatos criminais, evidenciando o comportamento do modelo na detecção desse tipo de entidade após a etapa de pseudo-rotulagem.

Tabela V.6: Exemplos de *tokens* anotados no formato BIO para a categoria Organização após o processo de ajuste fino do modelo. A tabela inclui siglas, nomes institucionais, termos relacionados a facções e plataformas digitais, demonstrando a capacidade do modelo em reconhecer diferentes tipos de organizações presentes nos relatos criminais.

Exemplos	Marcação
53dp	B-ORGANIZAÇÃO
facção	B-ORGANIZAÇÃO
cv	I-ORGANIZAÇÃO
corregedoria	B-ORGANIZAÇÃO
whatsapp	B-ORGANIZAÇÃO

V.3 Configurações Experimentais

Os experimentos foram executados em diferentes servidores de alto desempenho, cada um equipado com configurações distintas de GPU, memória e capacidade de processamento. As seguintes infraestruturas foram utilizadas:

- Servidor com GPU **RTX A6000**, 50 GB de RAM e 9 vCPUs;
- Servidor com GPU **NVIDIA A100 PCIe**, 80 GB de VRAM, 117 GB de RAM e 8 vCPUs;
- Servidor com GPU **NVIDIA RTX A100**, 80 GB de VRAM, 170 GB de RAM e 12 vCPUs.

Todos os experimentos são conduzidos utilizando um conjunto de bibliotecas Python com versões específicas, de modo a assegurar consistência, reprodutibilidade e alinhamento com a metodologia proposta. As principais bibliotecas empregadas foram:

- **Gliner** – versão 0.2.13;
- **PyTorch** – versão 2.5.0;
- **Matplotlib** – versão 3.10.1;
- **NumPy** – versão 2.1.2;
- **Scikit-learn** – versão 1.5.2;
- **Pandas** – versão 2.2.3.

Para o ajuste fino do modelo GLiNER, são adotados os seguintes hiperparâmetros: taxa de aprendizado de 3×10^{-5}, decaimento de pesos de 0,01, tamanho de lote de 8 exemplos e semente aleatória fixada em 42. A utilização de uma semente fixa tem como objetivo garantir a reprodutibilidade dos experimentos, enquanto os valores de taxa de aprendizado e decaimento de pesos

são definidos a partir de testes preliminares realizados durante a etapa de validação cruzada. O treinamento é conduzido utilizando mini-lotes de oito exemplos por iteração, buscando equilibrar o consumo de memória das GPUs disponíveis e a estabilidade do processo de otimização.

A definição explícita dessas versões assegura que todo o ambiente experimental possa ser reproduzido integralmente, preservando a confiabilidade dos resultados obtidos.

V.4 Comparando alternativas do modelo GliNER

Os modelos foram treinados com ajuste fino utilizando a base DD_{small}, conforme descrito na Seção IV.3.4. Inicialmente, o conjunto de dados é dividido em 20% para teste e 80% para treino. A porção destinada ao treino é particionada em cinco subconjuntos de igual tamanho, permitindo a aplicação da validação cruzada do tipo *5-fold*. Em cada *fold*, são realizadas cinco execuções independentes, possibilitando a avaliação da estabilidade do modelo frente a diferentes inicializações e subconjuntos de dados. A Tabela V.7 apresenta os resultados obtidos em cada etapa do processo, incluindo os valores de *F1* por partição, a definição dos limiares de confiança na validação e o desempenho final no conjunto de teste.

Em seguida, é realizada a etapa de validação em cada *fold*, na qual são avaliados diferentes limiares de confiança com o objetivo de identificar a configuração que apresentasse o melhor desempenho. Após a definição do limiar ótimo, o modelo correspondente é avaliado no conjunto de teste previamente separado, sem a aplicação do teste de sanidade, cenário em que há maior probabilidade de repetição de entidades já observadas durante o treinamento. Posteriormente, é aplicado o teste de sanidade, com o propósito de avaliar de forma mais rigorosa a capacidade de generalização do modelo, considerando apenas entidades não vistas anteriormente no processo de treino.

Para diferenciar o desempenho do modelo GLiNER (Seção II.2.3), são avaliados diversos hiperparâmetros, incluindo a taxa de aprendizado (*learning rate*) e o decaimento de pesos (*weight decay*). Além disso, diferentes valores de limiar de confiança (*threshold*) são testados para analisar o comportamento do modelo em distintos cenários de pontuação.

Com base nos resultados obtidos pela validação cruzada, conforme ilustrado na Figura IV.3, seleciona-se a configuração de hiperparâmetros com o melhor desempenho médio. O *fold* associado ao maior valor de *F1* é então utilizado na avaliação final sobre o conjunto de teste previamente separado, alcançando um *F1-Score* de 78,21%. É importante destacar que esse resultado é obtido **sem a aplicação do teste de sanidade**, ou seja, em um cenário no qual o conjunto de teste ainda pode conter entidades previamente observadas durante o treinamento. Dessa forma, o valor de 78,21% representa o desempenho do modelo na configuração padrão de avaliação, enquanto análises mais rigorosas de generalização, baseadas exclusivamente em entidades inéditas, são apresentadas posteriormente por meio do teste de sanidade. Esse procedimento permite diferenciar o desempenho

Tabela V.7: Resultados dos testes para o modelo GliNER, considerando o *corpus* DD_{small} dividido em 5 *folds*. Após a validação cruzada externa, foi realizada validação interna com diferentes limiares de decisão (0,5 e 0,6). O melhor modelo foi selecionado com base no desempenho obtido e posteriormente avaliado no conjunto de teste separado, alcançando F1 de 78,21%.

Etapa	Fold	Limiar de Confiança	F1
Treino	Fold 1	–	41,03%
Treino	Fold 2	–	41,82%
Treino	Fold 3	–	41,22%
Treino	Fold 4	–	41,49%
Treino	Fold 5	–	41,14%
Validação	Fold 4	0,5	41,29%
Validação	Fold 4	0,6	41,38%
Validação	Melhor Modelo (Fold 4)	0,6	**78,21%**
Teste	Teste de Sanidade	0,6	60,58%

em condições convencionais daquele observado em um cenário mais desafiador de generalização.

Ao final dos experimentos realizados sobre o conjunto DD_{small}, é aplicado um teste de sanidade com o objetivo de avaliar o desempenho do modelo em entidades que não haviam sido utilizadas no processo de treino. Esse procedimento produz um *F1* de 45,31%, servindo como base para a análise da robustez e capacidade de generalização dos modelos desenvolvidos.

V.5 Comparação com outros modelos

Os modelos selecionados para as comparações são escolhidos com base nos trabalhos analisados na Seção III.3.1, que fundamentam esta dissertação. Entre eles, destacam-se sistemas amplamente consolidados e utilizados na literatura recente, como o GPT e o Gemini, ambos representativos da abordagem baseada em LLMs.

Além desses, considerada-se o Spacy, um modelo mais leve e tradicional, que não se baseia em LLMs, mas é amplamente empregado em aplicações práticas de processamento de linguagem natural. Para fins de comparação com arquiteturas mais próximas ao GLiNER, foi seleciona-se o BERTimbau, desenvolvido inicialmente por Souza et al. [2020], que utiliza a biblioteca *Transformers* como tecnologia central e apresenta resultados relevantes em tarefas de reconhecimento de entidades em língua portuguesa.

Como referência adicional, inclui-se ainda o modelo baseado em LSTM, conforme apresentado em Figueiredo et al. [2023], o qual serve como ponto de partida para as análises iniciais deste trabalho e possibilita uma comparação entre arquiteturas recorrentes tradicionais e abordagens

mais recentes baseadas em mecanismos de atenção.

É importante ressaltar que os modelos utilizados como *benchmark* não são avaliados por meio de validação cruzada com *folds*, diferentemente do GLiNER. Para esses modelos, a avaliação é realizada diretamente sobre o conjunto de teste utilizado na comparação final. Dessa forma, os resultados de GPT, Gemini, Spacy, BERTimbau e LSTM refletem uma única execução experimental, enquanto os resultados do GLiNER são obtidos a partir do processo de seleção e validação descrito na Seção IV.3.4.

Conforme a Tabela V.8, os modelos avaliados não apresentaram resultados expressivos devido às particularidades dos *corpora* utilizados no processo de ajuste fino, bem como aos testes que exigiram capacidade de generalização. Observa-se, entretanto, que os melhores desempenhos são obtidos por modelos cuja finalidade principal consiste no reconhecimento de entidades, em contraste com modelos generalistas.

Tabela V.8: Comparação entre diferentes modelos aplicados à tarefa de reconhecimento de entidades, considerando as métricas de F1, Revocação e Precisão.

Modelo	Versão	Ajuste Fino	F1	Revocação	Precisão
GliNER	multilanguage	✓	0,60	0,64	0,65
GliNER	large	✓	0,58	0,60	0,64
LSTM	–	✓	0,28	0,14	0,13
Gemini	2.5-flash	✓	0,33	0,20	0,91
Gemini	2.5-flash	–	0,11	0,06	0,71
Gemini	1.5	–	0,02	0,01	0,02
Gemini	1.5-flash	–	0,14	0,10	0,22
GPT	4	✓	0,24	0,17	0,43
GPT	4	–	0,07	0,05	0,11
GPT	4-turbo	✓	0,24	0,17	0,40
GPT	3.5-turbo	✓	0,25	0,20	0,33
Spacy	pt-core-news-lg	✓	0,43	0,37	0,52
Bertimbau	bert-base-portuguese-cased	✓	0,56	0,52	0,62

Os modelos GLiNER apresentaram os melhores resultados gerais. A versão *multilanguage* obteve o maior *F1*, alcançado 0,60, com *Revocação* de 0,64 e Precisão de 0,65. A versão *large*, embora ligeiramente inferior, ainda apresentou forte desempenho, com *F1* de 0,58. Esses resultados evidenciam a robustez do GLiNER quando submetido ao ajuste fino, destacando-o como o modelo mais eficaz da comparação.

O BERTimbau, fundamentado na arquitetura *Transformers*, apresentou desempenho competitivo, conforme evidenciado no estudo de da Silva et al. [2023], alcançando *F1-Score* de 0,56, com

valores equilibrados de *Revocação* (0,52) e *Precisão* (0,62). Esses resultados demonstram sua adequação para tarefas de Reconhecimento de Entidades Nomeadas em língua portuguesa.

Em contraste, o modelo LSTM, mesmo com ajuste fino, apresentou desempenho significativamente inferior, obtendo *F1-Score* de 0,28. Esse resultado reforça as limitações dessa arquitetura em tarefas mais complexas de reconhecimento de entidades, especialmente quando comparada a modelos baseados em *Transformers*, que conseguem capturar melhor dependências contextuais de longo alcance.

O Spacy, utilizando o modelo *pt-core-news-lg*, apresentou desempenho intermediário, com *F1* de 0,43. Embora eficiente em aplicações práticas, o Spacy demonstra limitações frente a modelos baseados em LLMs, especialmente em cenários que exigem maior capacidade de generalização.

Por fim, os modelos GPT e Gemini apresentaram desempenhos inferiores, sobretudo nas versões não ajustadas. No caso do GPT, os valores de *F1* variaram de 0,07 (GPT-4 sem ajuste fino) e 0,25 (GPT-3.5-turbo ajustado), com baixos valores de *Revocação*. O Gemini apresentou resultados ainda mais modestos, variando entre 0,02 (versão 1.5) e 0,33 (versão Gemini 2.5-flash com ajuste fino), embora algumas versões tenham exibido alta precisão, evidenciando forte desequilíbrio nas métricas.

V.5.1 Prompts Utilizados nos Modelos GPT e Gemini

Para permitir uma comparação padronizada entre os modelos de linguagem de grande porte (LLMs), foram definidos *prompts* específicos para cada arquitetura, considerando os formatos de saída mais adequados às suas capacidades de processamento.

No caso do modelo GPT, adota-se uma estratégia baseada em rotulação sequencial de *tokens* utilizando o esquema BIO, compatível com a tarefa de REN. O *prompt* utilizado foi:

> Você é um modelo de NER. Sua tarefa é rotular cada token com BIO usando apenas: O, B-PER, I-PER, B-LOC, I-LOC, B-ORG, I-ORG.

Essa abordagem solicita que o modelo realize a classificação de cada *token* individualmente, indicando o início (`B`), a continuação (`I`) ou a ausência (`O`) de uma entidade nomeada pertencente às categorias Pessoa (`PER`), Localização (`LOC`) ou Organização (`ORG`).

Para o modelo Gemini, optou-se por um formato baseado em extração direta de entidades utilizando deslocamentos de caracteres, permitindo posterior conversão para o esquema de avaliação adotado. O *prompt* empregado foi:

> Você é um extrator de entidades nomeadas (NER). Extraia entidades exatas no texto. Use apenas as labels: Person, Location, Organization. Retorne estritamente JSON com este formato:
>
> `"results":["start":int,"end":int,"label":string]`

- Os índices são offsets de caractere em relação a `text` (0-based), no intervalo `[start, end)`.
- Não inclua comentários, markdown, explicações ou texto fora do JSON.
- Não invente entidades.
- Se não houver entidades, retorne exatamente: `"results":[]`.

A utilização de formatos distintos decorre das características de cada modelo. Enquanto o GPT apresentou melhor aderência à rotulação token a token no formato BIO, o Gemini produziu resultados mais consistentes quando instruído a retornar diretamente os limites das entidades identificadas em formato JSON estruturado.

V.6 Calibração da pseudo-rotulação do *corpus* DD_{large}

A calibração é realizada a partir de um conjunto de 100 relatos criminais selecionados aleatoriamente do *corpus* DD_{large}, previamente validados manualmente. Esse subconjunto contém 670 entidades, e é utilizado como base de referência para avaliar o desempenho do modelo que apresentou melhor performance nas etapas anteriores. A aplicação de técnicas de calibração torna-se necessária em virtude do comportamento de superconfiança observado, uma vez que o modelo atribuí valores de confiança iguais ou superiores a 0,9 à maioria das predições. Diante desse cenário, busca-se ajustar a distribuição das probabilidades estimadas, a fim de torná-las mais condizentes com a acurácia real do modelo e, consequentemente, permitir uma avaliação mais confiável sob diferentes abordagens de calibração.

A Figura V.1 ilustra a distribuição das entidades no *corpus* DD_{large} de acordo com os valores de confiança estimados pelo modelo sem calibração. Observa-se que a maior parte das entidades se concentra em valores próximos a 1, indicando uma clara tendência do modelo em atribuir níveis excessivamente elevados de confiança às suas predições.

Esse comportamento evidencia um viés de superconfiança, uma vez que a distribuição apresenta forte assimetria e baixa uniformidade ao longo das faixas de pontuação. Embora existam entidades distribuídas entre os intervalos de 0 a 0,9, a frequência nessas regiões é significativamente menor quando comparada à região de 0,9 a 1,0, onde ocorre um pico acentuado.

Tal resultado justifica a adoção de técnicas de calibração, como *Temperature Scaling* e *Isotonic Regression*, que permitem ajustar a confiabilidade dos valores atribuídos e reduzir o risco de interpretações equivocadas sobre a segurança do modelo em suas decisões.

A Figura V.2 apresenta a distribuição dos valores de confiança após a aplicação da técnica *Temperature Scaling* sobre o *corpus* DD_{large}. Diferentemente do cenário sem calibração, observa-se maior dispersão ao longo das faixas de score, com concentrações relevantes entre 0,4 e 0,7, além de um pico expressivo próximo a 0,9.

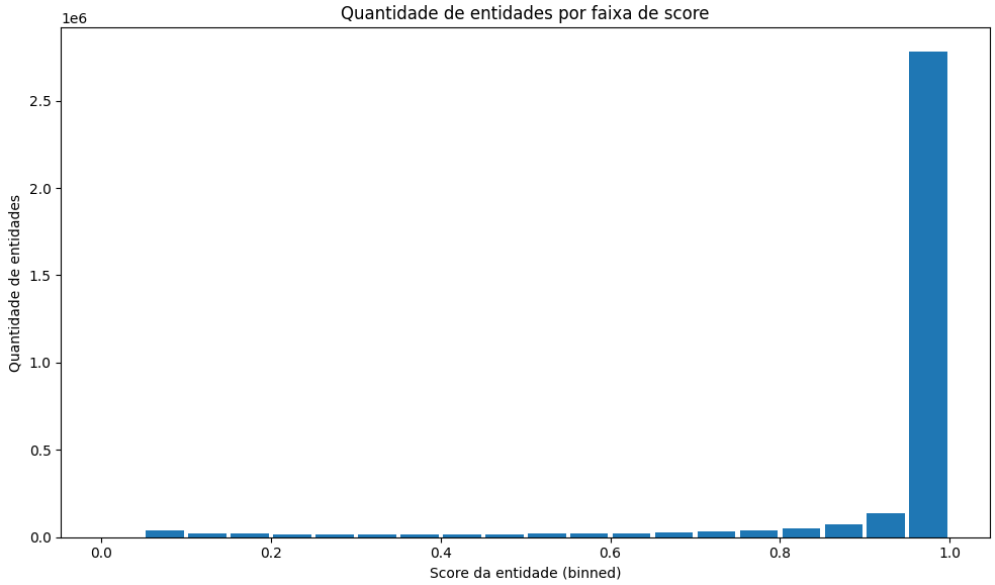


Figura V.1: Frequência da confiança das entidades no *corpus* DD_{large}, mostrando a distribuição das entidades por faixas de score de confiança atribuídos pelo modelo.

Esse comportamento demonstra que a calibração contribuiu para atenuar a superconfiança do modelo, redistribuindo os valores previstos de forma mais equilibrada. A presença de diferentes faixas de concentração indica que o *Temperature Scaling* ajustou adequadamente a escala dos valores de confiança, tornando-os mais compatíveis com o grau real de incerteza das predições. Assim, o modelo calibrado apresenta maior robustez e oferece valores de confiança mais interpretáveis.

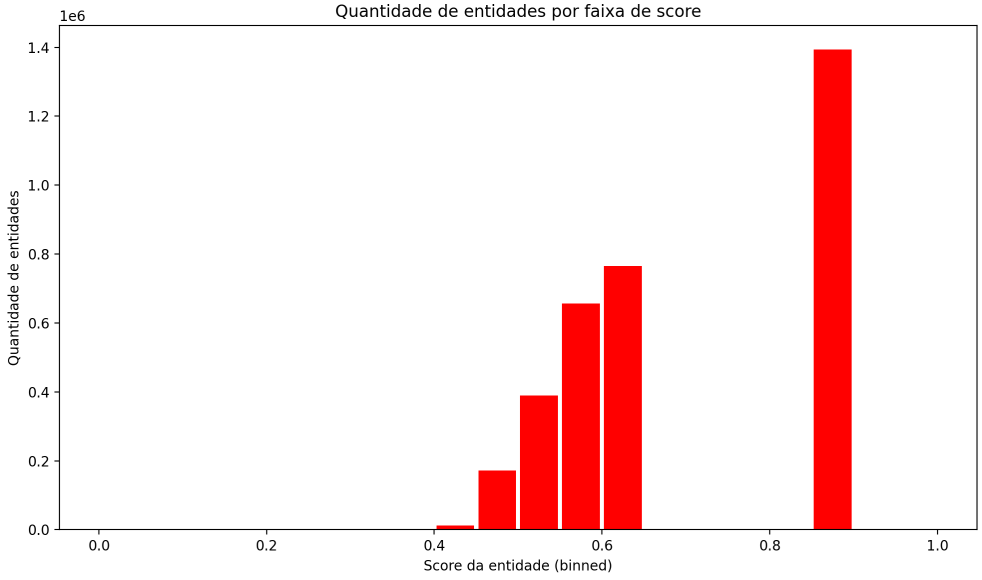


Figura V.2: Distribuição dos valores de confiança no *corpus* DD_{large} após a aplicação da técnica *Temperature Scaling*, evidenciando a maior dispersão dos escores calibrados.

A Figura V.3 apresenta a distribuição dos valores de confiança após a aplicação da técnica *Isotonic Regression* no *corpus* DD_{large}. Nesse caso, a distribuição torna-se ainda mais dispersa,

com concentrações distribuídas em diferentes intervalos de score, entre as faixas de 0,3 e 0,8.

Esse padrão indica que o método *Isotonic Regression* suaviza de maneira mais incisiva a tendência de superconfiança do modelo, redistribuindo os valores previstos em um espectro mais amplo. Como consequência, há uma representação mais fiel da incerteza associada às predições, permitindo interpretações mais consistentes. Assim, a técnica mostra-se eficaz na correção do viés inicial do modelo, ajustando os escores a uma escala coerente com a performance real observada.

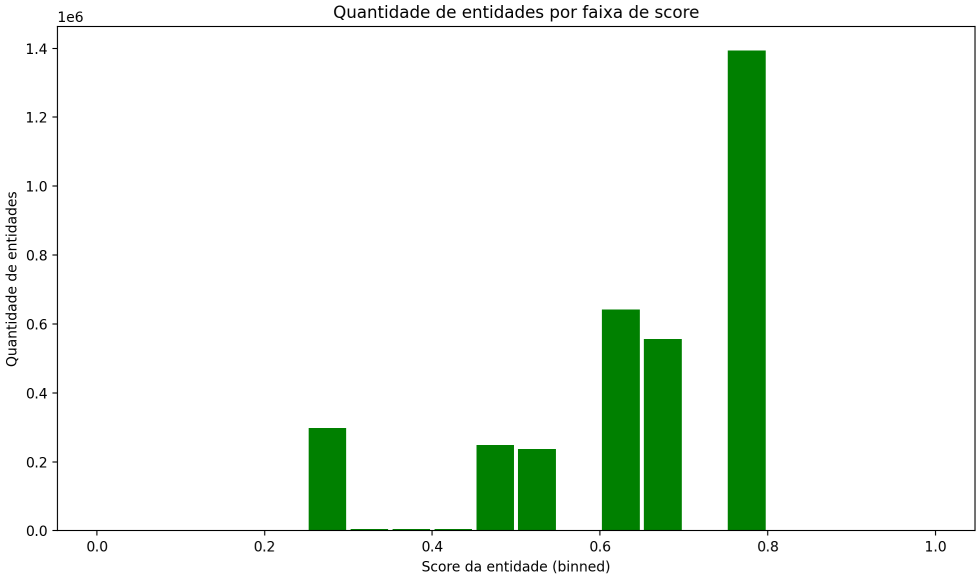


Figura V.3: Distribuição dos valores de confiança no *corpus* DD_{large} após a aplicação da técnica *Isotonic Regression*, evidenciando maior uniformidade e suavização dos escores.

V.6.1 Métricas da Calibração

Curvas de Confiabilidade

O gráfico da Figura V.4 compara o grau de calibração dos modelos em três condições: sem calibração (RAW), com *Temperature Scaling* (TS) e com *Isotonic Regression* (ISO). A linha pontilhada indica a curva ideal, que representa o alinhamento perfeito entre confiança prevista e acurácia empírica.

A curva do modelo sem calibração apresenta desvio expressivo em relação à curva ideal, especialmente na faixa entre 0,6 e 0,8 de confiança média. Nesses intervalos, a acurácia obtida é significativamente inferior à confiança declarada, evidenciando o comportamento de superconfiança. Esse resultado reforça a necessidade de aplicar técnicas de calibração para reduzir o descompasso entre confiança e desempenho real.

O modelo calibrado com *Temperature Scaling* demonstra maior proximidade da curva ideal, sobretudo para valores altos de confiança. Embora ainda existam discrepâncias, a técnica corrige de forma relevante a superconfiança observada, produzindo uma relação mais coerente entre confiança

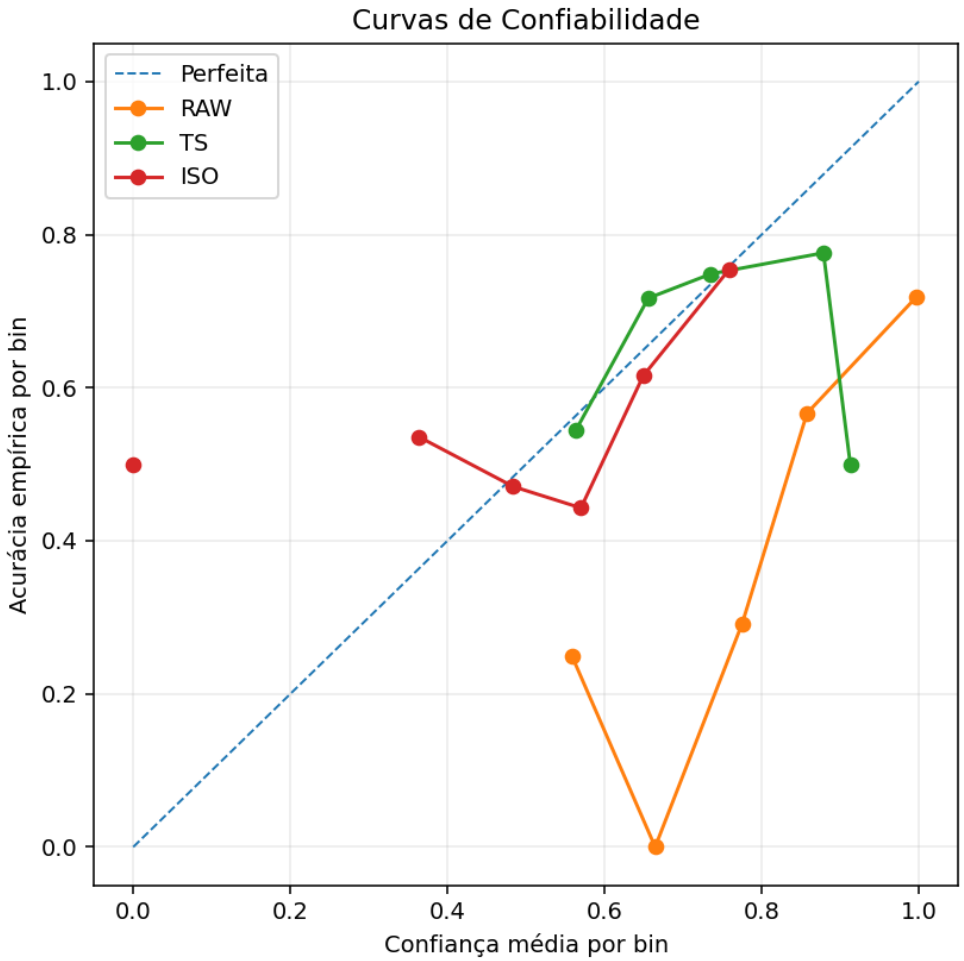


Figura V.4: Curvas de confiabilidade para os métodos RAW (**Sem Calibração**), TS (**Temperature Scaling**) e ISO (**Isolation Regression**). A linha tracejada diagonal representa a calibração perfeita, na qual a confiança prevista pelo modelo é igual à proporção real de acertos observada em cada intervalo de confiança. Quanto mais próxima a curva estiver dessa linha, melhor calibradas são as estimativas de confiança do modelo.

média e acurácia empírica.

Já o modelo calibrado com *Isotonic Regression* apresenta a curva mais estável entre as três abordagens testadas. A técnica suaviza de maneira mais eficiente as inconsistências, distribuindo as estimativas de confiança de forma mais consistente. Em diversos trechos, a curva permanece próxima à ideal, indicando que o método é particularmente eficaz na correção do desalinhamento entre confiança e acurácia.

Métricas da Calibração

A Tabela V.9 apresenta as principais métricas de avaliação obtidas após a aplicação das técnicas de calibração, considerando um conjunto de 100 relatos criminais validados manualmente do *corpus* DD_{large}. As métricas permitem comparar o desempenho do modelo sem calibração e calibrado pelas técnicas *Temperature Scaling* e *Isotonic Regression*, oferecendo uma análise mais precisa sobre o

Tabela V.9: Resultados das métricas de calibração considerando ECE, MCE, Brier Score, NLL e Acurácia para os modelos sem calibração, calibrado com *Isotonic Regression* e calibrado com *Temperature Scaling*.

Base de Dados	ECE	MCE	Brier	NLL	Acurácia
Sem Calibração	0,28	0,67	0,28	2,97	0,57
Isotonic Regression	0,06	0,36	0,20	0,67	0,59
Temperature Scaling	0,07	0,20	0,21	0,61	0,57

impacto dessas abordagens na confiabilidade das predições.

Os resultados indicam que a acurácia geral permanece semelhante entre os três modelos, o que demonstra que o objetivo da calibração não é melhorar a taxa de acerto, mas sim ajustar a relação entre confiança predita e acurácia observada. Esse comportamento é esperado, dado que as técnicas de calibração atuam na distribuição dos escores de confiança e não diretamente na capacidade discriminativa do modelo.

Ao analisar a métrica ECE, nota-se uma redução substancial nos modelos calibrados, especialmente no caso da *Isotonic Regression*. O modelo sem calibração apresenta um ECE elevado, indicando superconfiança pronunciada. Tanto o *Temperature Scaling* quanto o *Isotonic Regression* melhoram significativamente esse cenário, tornando a distribuição das probabilidades mais fiel ao desempenho real.

A métrica *Maximum Calibration Error* (MCE) reforça essas observações. O modelo sem calibração apresenta discrepâncias máximas elevadas, enquanto o *Temperature Scaling* reduz esse efeito e a *Isotonic Regression* apresenta o melhor desempenho geral, suavizando os desvios mais extremos.

O *Brier Score* também confirma o impacto das técnicas de calibração na qualidade probabilística das predições. A calibração reduz o erro médio das probabilidades atribuídas, com destaque para o *Isotonic Regression*, que obtém o valor mais baixo entre as três abordagens, indicando uma distribuição mais confiável e informativa dos scores.

Por fim, a métrica NLL evidencia o desalinhamento inicial do modelo sem calibração, que apresenta valores elevados, caracterizando forte superconfiança. O *Temperature Scaling* reduz expressivamente esse valor, aproximando-se de um cenário ideal. Embora o *Isotonic Regression* também melhore a métrica em relação ao modelo não calibrado, o *Temperature Scaling* apresenta desempenho ligeiramente superior em NLL. No conjunto, os resultados demonstram que ambas as técnicas são eficazes na mitigação da superconfiança, produzindo previsões mais consistentes e calibradas.

V.7 Impacto da pseudo-rotulação do *corpus* $DD_{\textbf{large}}$

Após a realização dos testes e das avaliações descritas nas seções anteriores, o modelo GLiNER foi aplicado ao *corpus* DD_{large}, que reúne um volume expressivo de relatos criminais, com o objetivo de identificar automaticamente as entidades nele descritas. Essa etapa visa ampliar o conjunto de dados anotados por meio da pseudo-rotulação, promovendo ganho quantitativo e potencial melhoria de reconhecimento das entidades nos *corpora*, conforme discutido no trabalho de Gomes Jr. et al. [2023].

A Tabela IV.4 apresenta o número total de relatos processados, bem como a quantidade de palavras presentes no corpus, fornecendo uma visão geral da escala dos dados utilizados nessa etapa.

Para a realização da pseudo-rotulação, emprega-se o algoritmo descrito no Algoritmo 1, no qual cada iteração do processo é analisada considerando três grupos distintos. Os experimentos consideraram tanto cenários sem calibração quanto configurações com a aplicação de técnicas de calibração, especificamente o *Temperature Scaling* e a *Isotonic Regression*. Além disso, avalia-se o impacto da variação dos valores de confianças associados aos metadados presentes no corpus DD_{large}, bem como cenários sem essa variação, garantindo uma análise comparativa abrangente.

Cada processo de ajuste fino foi configurado para até 10 épocas de treinamento. No entanto, adota-se a técnica de *Early Stopping*, estabelecendo que, caso a principal métrica de avaliação, o *F1*, não apresente melhora ao longo de três épocas consecutivas, o treinamento seja interrompido antecipadamente e o modelo correspondente seja salvo como a melhor versão até então.

Para garantir a comparabilidade dos resultados entre as diferentes configurações experimentais, todos os modelos gerados ao longo do processo de pseudo-rotulação são avaliados sobre o mesmo conjunto de teste definido na Figura IV.2, correspondente a $DD^{\text{teste}}_{\text{small}}$. Esse subconjunto permanece reservado exclusivamente para avaliação final, não sendo utilizado em nenhuma etapa de treinamento, calibração ou geração de pseudo-rótulos. Assim, as métricas reportadas refletem a capacidade de generalização dos modelos diante de dados previamente não observados.

Os experimentos apresentados nesta seção têm como objetivo analisar a influência dos metadados estruturados no processo de pseudo-rotulação. Em particular, busca-se responder às questões de pesquisa relacionadas ao impacto do uso de metadados e à sua interação com técnicas de calibração, avaliando se essas informações adicionais contribuem para melhorar o desempenho do modelo na tarefa de reconhecimento de entidades nomeadas.

V.7.1 Utilização de Metadados

Para a utilização dos metadados, aplica-se uma variação de 20% adicional sobre os valores de confiança atribuídos pelo modelo, respeitando o limite máximo de 1. A partir dessa estratégia, diferentes níveis de confiança são definidos e avaliados para determinar a quantidade de relatos criminais selecionados para compor o conjunto de treinamento durante o ajuste fino, considerando distintos volumes de relatos.

Inicialmente, avalia-se o cenário em que os metadados são incorporados ao processo de seleção dos relatos, porém sem a aplicação de técnicas de calibração. Essa análise permite verificar se o uso isolado dos metadados é capaz de produzir ganhos de desempenho em relação aos cenários que não utilizam informações estruturadas.

Sem Calibração

A Tabela V.10 apresenta os resultados das iterações de pseudo-rotulação realizadas sem a aplicação de técnicas de calibração, utilizando diferentes limiares de confiança dos relatos presentes no *corpus* DD_{large} para a seleção dos dados.

Na primeira iteração, utilizando o limiar de confiança de 0,9, foram mantidos 298.614 relatos criminais e descartados 68.438, resultando em um *F1* de 65,4%, com precisão de 65,9% e *revocação* de 65%. Na segunda iteração, com limiar 0,8, foram mantidos 60.296 relatos e descartados 8.142, apresentando leve melhora no desempenho, com *F1* de 66,3%.

Na terceira iteração, com limiar de 0,7, foram mantidos apenas 3.582 relatos e descartados 4.560; porém, com desempenho similar ao da segunda iteração: *F1* de 66,3%, precisão de 66,5% e *revocação* de 66,2%. A quarta iteração, com limiar de 0,6, manteve apenas 316 relatos e descartou 4.244, alcançando o melhor desempenho do grupo, com *F1* de 67,0% e o maior valor de *revocação*.

Por fim, a quinta iteração, com limiar 0,1, selecionou 664 relatos e descartou 3.580, mas apresentou queda no desempenho, atingindo *F1* de 61,3%, precisão de 59,6% e *revocação* de 63,0%.

Tabela V.10: Resultados das iterações de pseudo-rotulação sem aplicação de técnicas de calibração, considerando diferentes níveis de confiança. A tabela apresenta a quantidade de relatos criminais mantidos e descartados em cada iteração, bem como as métricas de desempenho (F1, precisão e revocação).

Iteração	Limiar	Mantidos	Descartados	F1	Precisão	Revocação
1	0,9	298.614	68.438	65,4%	65,9%	65,0%
2	0,8	60.296	8.142	66,3%	66,5%	66,2%
3	0,7	3.582	4.560	66,3%	66,5%	66,2%
4	0,6	316	4.244	67,0%	66,2%	67,9%
5	0,1	664	3.580	61,3%	59,6%	63,0%

A Figura V.5 apresenta a matriz de confusão correspondente ao melhor resultado obtido na quarta iteração sem calibração. Observa-se que a classe **Localização** concentra a maior quantidade de predições corretas, com 945 ocorrências corretamente classificadas. Entretanto, ainda são observadas confusões entre as classes **Pessoa** e **Localização**, evidenciadas por 130 instâncias em que entidades de pessoa foram classificadas como Localização. Esse comportamento indica que determinadas expressões utilizadas nos relatos criminais podem assumir características ambíguas entre indivíduos e locais, dificultando a distinção pelo modelo.

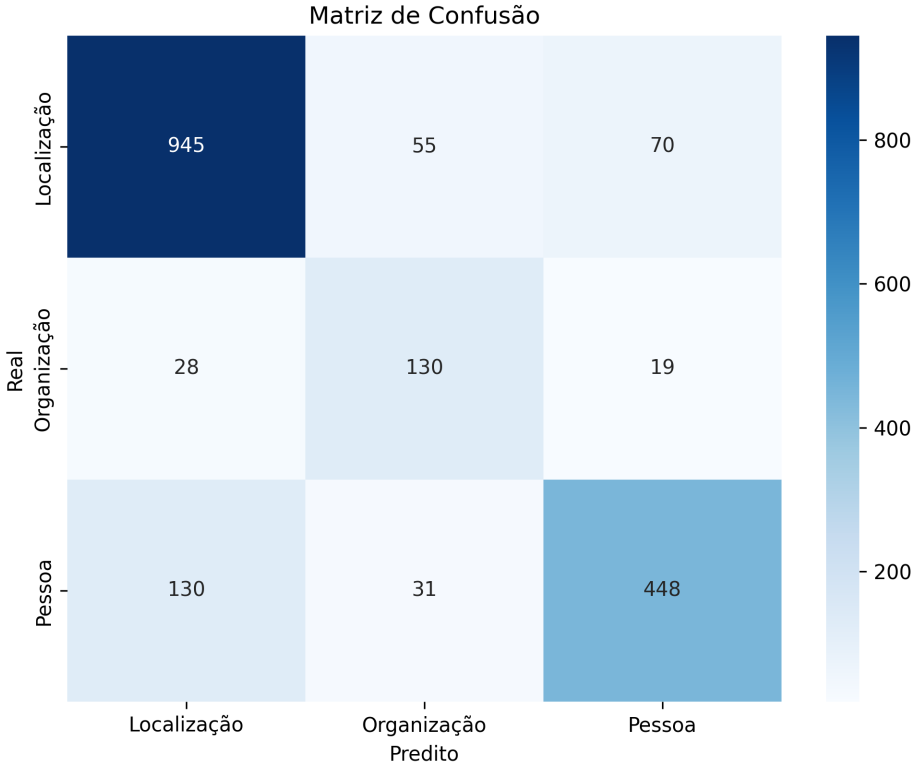


Figura V.5: Matriz de confusão obtida na avaliação do modelo sem aplicação de técnicas de calibração na quarta iteração disponível. A figura apresenta a distribuição das predições corretas e incorretas entre as classes **Localização**, **Organização** e **Pessoa**.

Em seguida, investiga-se o efeito combinado entre metadados e calibração dos escores de confiança. O objetivo é verificar se a redução da superconfiança do modelo, associada ao uso das informações estruturadas disponíveis no corpus, contribui para uma seleção mais eficiente dos relatos pseudo-rotulados e para a obtenção de melhores resultados.

Com Calibração

A Tabela V.11 apresenta os resultados obtidos com a aplicação da técnica de *Temperature Scaling* durante o processo de pseudo-rotulação. Foram realizadas três iterações, cada uma adotando diferentes limiares de confiança.

Na primeira iteração, com limiar de 0,9, foram mantidos 34.244 relatos e descartados 332.610,

alcançando um *F1* de 65,3%, com precisão de 65,9% e *revocação* de 64,7%. Na segunda iteração, utilizando um limiar de 0,8, foram mantidos apenas 1.943 relatos criminais e descartados 330.667, o que resultou em queda no desempenho geral e um *F1* de 58,7%.

Na terceira iteração, foram aceitos todos os relatos anteriormente descartados (limiar = 0), totalizando 330.667 relatos mantidos, sem descartes. O desempenho obtido foi *F1* de 62,2%, precisão de 60,67% e *revocação* de 63,8%.

Tabela V.11: Resultados das iterações de pseudo-rotulação com aplicação da técnica de calibração *Temperature Scaling*, considerando diferentes níveis de confiança.

Iteração	Limiar	Mantidos	Descartados	F1	Precisão	Revocação
1	0,9	34.244	332.610	65,3%	65,9%	64,7%
2	0,8	1.943	330.667	58,7%	52,5%	66,2%
3	0,0	330.667	0	62,2%	60,7%	63,8%

A Figura V.6 apresenta a matriz de confusão obtida após a aplicação da técnica de calibração *Temperature Scaling*. Observa-se que a classe **Localização** apresentou o maior número de predições corretas, com 1.021 ocorrências classificadas corretamente. A classe **Organização** também apresentou desempenho consistente, com 120 classificações corretas e baixo número de confusões com outras categorias.

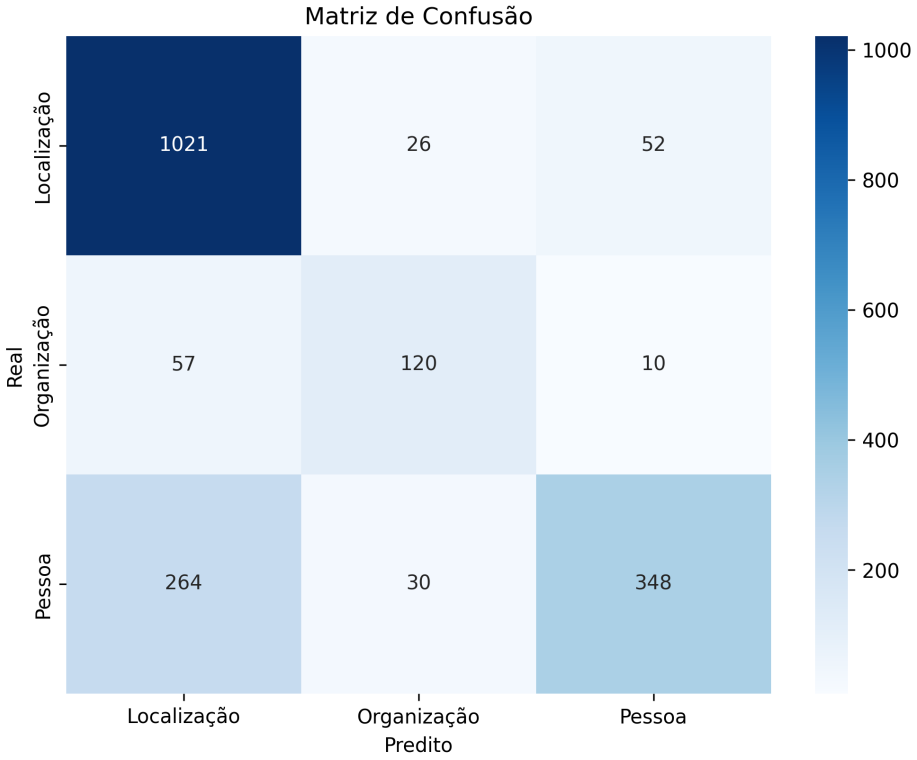


Figura V.6: Matriz de confusão obtida após a aplicação da técnica de calibração *Temperature Scaling* na primeira iteração disponível. A figura apresenta a distribuição das predições corretas e incorretas entre as classes **Localização**, **Organização** e **Pessoa**.

A Tabela V.12 apresenta os resultados referentes à utilização da técnica de *Isotonic Regression* na pseudo-rotulação. Os experimentos consideraram diferentes limiares de confiança para análise comparativa.

Na primeira iteração, com limiar 0,9, foram mantidos 11.672 relatos e descartados 355.182, obtendo *F1* de 62,3%, precisão de 65,9% e *revocação* de 59,1%. Na segunda iteração, com limiar 0,8, o número de relatos mantidos aumentou para 94.287, enquanto 260.895 foram descartados, apresentando desempenho semelhante (*F1* de 62,1%).

Na terceira iteração, com limiar 0,7, foram mantidos 164.573 relatos e descartados 96.322. Apesar da queda no *F1* para 61,6%, houve aumento da precisão (67%) e redução do *revocação* (57%). Por fim, na quarta iteração, todos os relatos foram mantidos (limiar 0), totalizando 96.322 relatos adicionais, alcançando *F1* de 67%, precisão de 66,2% e *revocação* de 67,9%.

Tabela V.12: Resultados das iterações de pseudo-rotulação com aplicação da técnica de calibração *Isotonic Regression*, considerando diferentes níveis de confiança e utilizando metadados. A tabela apresenta a quantidade de relatos criminais mantidos e descartados em cada iteração, bem como as métricas de desempenho (F1, precisão e revocação).

Iteração	Limiar	Mantidos	Descartados	F1	Precisão	Revocação
1	0,9	11.672	355.182	62,3%	65,9%	59,1%
2	0,8	94.287	260.895	62,1%	66,4%	58,3%
3	0,7	164.573	96.322	61,6%	67,0%	57,0%
4	0,0	96.322	0	**67,0%**	66,2%	67,9%

A Figura V.7 apresenta a matriz de confusão obtida após a aplicação da técnica de calibração *Isotonic Regression*. Observa-se que a classe **Localização** apresentou o maior número de classificações corretas, com 913 ocorrências corretamente identificadas. A classe **Organização** também apresentou desempenho consistente, com 134 predições corretas e baixo número de confusões com as demais categorias.

Após a análise dos cenários que incorporam informações estruturadas, são avaliados os experimentos sem a utilização de metadados. Essa comparação permite isolar a contribuição dessas informações adicionais e verificar em que medida os resultados obtidos dependem exclusivamente das predições geradas pelo modelo.

V.7.2 Sem Utilizar os metadados

Nesta seção, os metadados não foram considerados para o acréscimo dos valores de confiança atribuídos pelo modelo. Assim, todas as iterações de pseudo-rotulação foram conduzidas exclusivamente com as predições originais, sem qualquer influência das informações adicionais presentes no *corpus* DD_{large}. Essa configuração permite avaliar o desempenho do modelo de forma indepen-

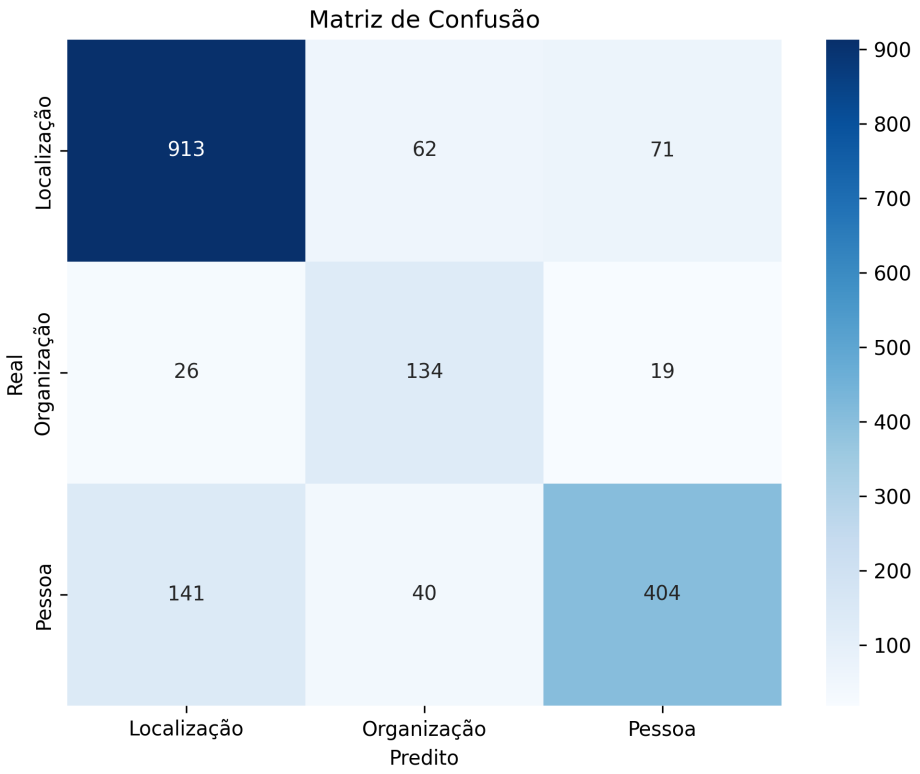


Figura V.7: Matriz de confusão obtida após a aplicação da técnica de calibração *Isotonic Regression* na quarta iteração disponível. A figura apresenta a distribuição das predições corretas e incorretas entre as classes **Localização**, **Organização** e **Pessoa**.

dente, isolando o impacto dos metadados e possibilitando uma análise comparativa direta entre cenários com e sem o uso desses recursos complementares.

Nesta etapa, o processo de pseudo-rotulação é realizado exclusivamente com os valores de confiança originalmente produzidos pelo modelo, sem o uso de metadados e sem qualquer procedimento de calibração. Os resultados obtidos servem como referência para comparação com os demais cenários experimentais.

Sem calibração

A Tabela V.13 apresenta os resultados das iterações de pseudo-rotulação realizadas sem a utilização de metadados e sem a aplicação de técnicas de calibração, considerando diferentes limiares de confiança para a seleção do *corpus* DD_{large}.

Na primeira iteração, utilizando o limiar de confiança em 0,9, foram mantidos 298.416 relatos e descartados 68.438. Nesse cenário, o modelo alcançou um *F1* de 64,1%, com precisão de 64,6% e *revocação* de 63,6%.

Na segunda iteração, com o limiar de confiança de 0,8, foram mantidos 60.706 e descartados 7.732. Observou-se uma melhora no desempenho, alcançando um *F1* de 65,4%, precisão de 66,6% e *revocação* de 64,2%. Esse resultado sugere que, mesmo na ausência de metadados e calibração,

limiares intermediários podem favorecer a qualidade do processo de pseudo-rotulação.

Na terceira iteração, utilizando o limiar de 0,7, o número de relatos mantidos reduziu-se para 3.249, enquanto 4.483 foram descartados. Nesse caso, houve leve queda no desempenho, com *F1* de 63,7%, precisão de 67,0% e *revocação* de 60,6%.

Por fim, na quarta iteração, correspondente ao limiar mínimo, foram mantidos 4.483 relatos, sem descartes adicionais. No entanto, o desempenho foi inferior ao das iterações anteriores, resultando em *F1* de 63,4%, precisão de 62,0% e *revocação* de 64,8%. Esse comportamento indica que a ausência de filtro por limiar de confiança não melhora a qualidade da pseudo-rotulação e pode, inclusive, comprometer a consistência das predições.

Tabela V.13: Resultados das iterações de pseudo-rotulação sem a utilização de metadados e sem aplicação de técnicas de calibração, considerando diferentes limiares de confiança. A tabela apresenta a quantidade de relatos criminais mantidos e descartados em cada iteração, além das métricas de desempenho (F1, precisão e revocação).

Iteração	Limiar	Mantidos	Descartados	F1	Precisão	Revocação
1	0,9	298.416	68.438	64,1%	64,6%	63,6%
2	0,8	60.706	7.732	65,4%	66,6%	64,2%
3	0,7	3.249	4.483	63,7%	67,0%	60,6%
4	0,0	4.483	0	63,4%	62,0%	64,8%

A Figura V.8 apresenta a matriz de confusão correspondente ao melhor resultado obtido nas iterações sem utilização de metadados e sem aplicação de técnicas de calibração. Observa-se que a classe *Localização* apresentou o maior número de classificações corretas, com 892 ocorrências identificadas corretamente. A classe *Organização* apresentou 118 classificações corretas, com poucas confusões com as demais categorias.

Por fim, avalia-se o impacto das técnicas de calibração quando aplicadas isoladamente, isto é, sem a utilização de metadados. Essa configuração permite verificar se a calibração dos escores de confiança é suficiente para melhorar a qualidade da seleção dos relatos e o desempenho final do modelo.

Com Calibração

A Tabela V.14 apresenta os resultados das iterações de pseudo-rotulação realizadas sem a utilização de metadados, mas com a aplicação da técnica de calibração *Temperature Scaling*, considerando diferentes limiares de confiança para a seleção dos relatos.

Na primeira iteração, utilizando um limiar de confiança de 0,8, foram mantidos 122.477 relatos e descartados 244.377. O modelo alcançou um *F1* de 62,4%, com precisão de 66,2% e *revocação* de 59,0%.

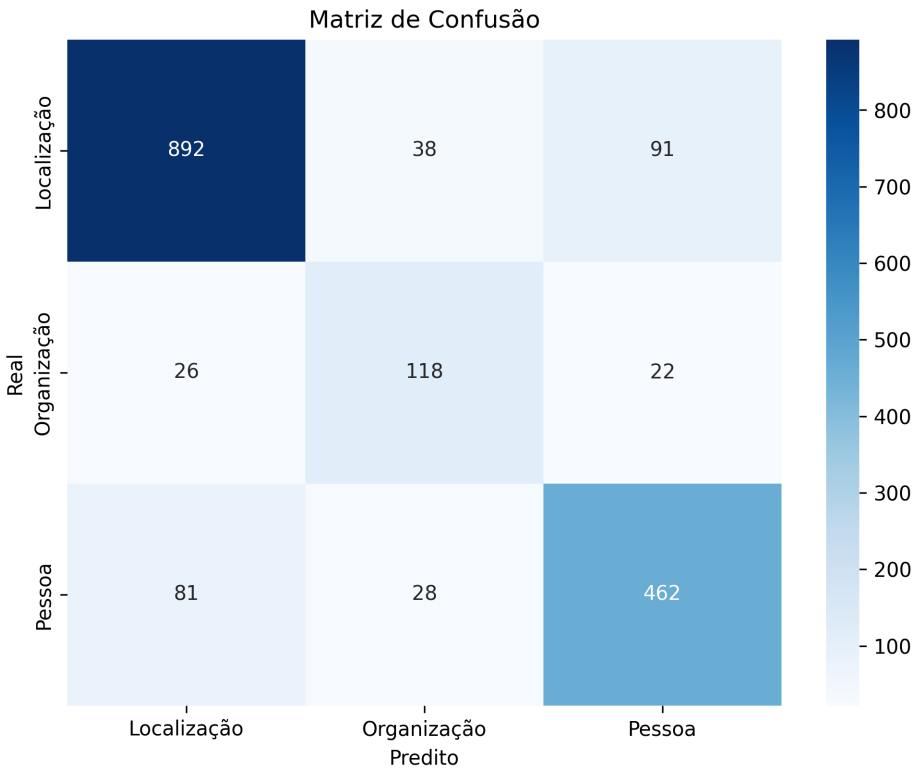


Figura V.8: Matriz de confusão obtida na avaliação do modelo treinado sem utilização de metadados e sem aplicação de técnicas de calibração na segunda iteração disponível. A figura apresenta a distribuição das predições corretas e incorretas entre as classes **Localização**, **Organização** e **Pessoa**.

Na segunda iteração, também com limiar de 0,8, mas aplicada em um novo ciclo de pseudo-rotulação, foram mantidos 116.875 relatos e descartados 127.502. O desempenho caiu significativamente, atingindo *F1* de 56,1%. A precisão manteve-se em 66,1%, mas o *revocação* reduziu-se para 48,7%, indicando perda de generalização.

Na terceira iteração, com limiar 0,7, foram mantidos 96.979 relatos e descartados 30.523. O desempenho continuou inferior ao anterior, com *F1* de 54,0%, precisão de 64,7% e *revocação* de 46,4%. Embora a precisão permaneça elevada, a queda acentuada no *revocação* compromete a cobertura das entidades.

Na quarta iteração, com limiar de 0,6, foram mantidos 22.240 relatos e descartados 8.283. Houve leve recuperação no *F1*, que atingiu 55,0%, com precisão de 64,2% e *revocação* de 48,1%. Ainda assim, a melhoria foi modesta, sugerindo que limiares mais baixos não garantem maior estabilidade.

Por fim, na quinta iteração, com limiar de confiança de 0,5, foram mantidos apenas 3.495 relatos e descartados 4.788. Nessa configuração, o modelo obteve *F1* de 57,9%, precisão de 62,2% e *revocação* de 54,1%. Embora o volume dos dados tenha sido reduzido, observou-se um equilíbrio ligeiramente melhor entre precisão e *revocação*, indicando que limiares menores podem recuperar parcialmente a performance, embora com risco de reduzir excessivamente o conjunto de treinamento.

Tabela V.14: Resultados das iterações de pseudo-rotulação realizadas sem a utilização de metadados, mas com aplicação da técnica de calibração *Temperature Scaling*. A tabela apresenta os valores correspondentes às iterações com diferentes limiares de confiança, incluindo a quantidade de relatos mantidos e descartados e as métricas de desempenho (F1, precisão e revocação).

Iteração	Limiar	Mantidos	Descartados	F1	Precisão	Revocação
1	0,8	122.477	244.377	62,4%	66,2%	59,0%
2	0,8	116.875	127.502	56,1%	66,1%	48,7%
3	0,7	96.979	30.523	54,0%	64,7%	46,4%
4	0,6	22.240	8.283	55,0%	64,2%	48,1%
5	0,5	3.495	4.788	57,9%	62,2%	54,1%

A Figura V.9 apresenta a matriz de confusão obtida na avaliação do modelo sem utilização de metadados, mas com aplicação da técnica de calibração *Temperature Scaling*. Observa-se que a classe **Localização** apresentou o maior número de classificações corretas, com 824 ocorrências corretamente identificadas. A classe *Organização* apresentou 121 classificações corretas e baixo número de confusões com as demais categorias.

A Tabela V.15 apresenta os resultados das iterações de pseudo-rotulação realizadas sem a utilização de metadados, mas com a aplicação da técnica de calibração *Isotonic Regression*, considerando diferentes limiares de confiança.

Na primeira iteração, com o limiar de 0,7, foram mantidos 151.354 relatos e descartados 215.500. O modelo alcançou um *F1* de 63,0%, com precisão de 68,4% e *revocação* de 58,4%.

Na segunda iteração com limiar de 0,8, foram mantidos 209.817 relatos e descartados 5.683. Contudo, o *F1* reduziu-se para 57,2%, influenciado pela queda no *revocação* para 49,6%, apesar da precisão elevada.

Na terceira iteração, com limiar 0,3, apenas 908 relatos foram mantidos, enquanto 4.775 foram descartados. Mesmo com o baixo volume de dados, o modelo apresentou *F1* de 58,0%, precisão de 67,6% e *revocação* de 50,8%.

Na quarta iteração, com limiar 0,2, foram mantidos 629 relatos e descartados 4.146. O modelo atingiu *F1* de 59,7%, com precisão de 58,0% e *revocação* de 61,5%, evidenciando maior equilíbrio entre as métricas.

Por fim, na quinta iteração, também com limiar 0,2, porém em um novo ciclo de pseudo-rotulação, foram mantidos 2.630 relatos e descartados 1.516. O desempenho resultante foi *F1* de 58,8%, precisão de 56,7% e *revocação* de 61,1%, reforçando a tendência de que limiares mais baixos favorecem maior cobertura.

De forma geral, a análise da Tabela V.15 evidencia que a aplicação da calibração por IR altera de maneira significativa a dinâmica entre precisão e revocação ao longo das iterações de pseudo-

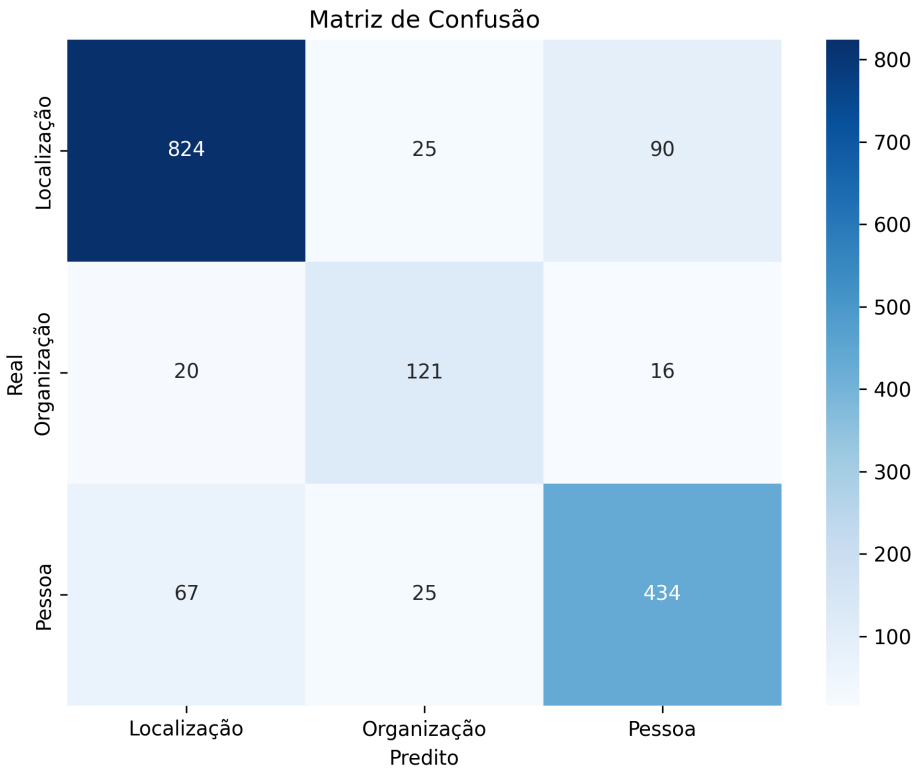


Figura V.9: Matriz de confusão obtida na avaliação do modelo treinado sem utilização de metadados, mas com aplicação da técnica de calibração *Temperature Scaling* na primeira iteração disponivel. A figura apresenta a distribuição das predições corretas e incorretas entre as classes **Localização**, **Organização** e **Pessoa**.

rotulação. Observa-se que limiares de confiança mais elevados conduzem a um comportamento mais conservador do modelo, priorizando a precisão (de no máximo 68,4%) em detrimento da revocação (com o valor de 58,4%), enquanto limiares mais baixos promovem maior cobertura das entidades, elevando a revocação (até 61,5%) às custas de uma redução na precisão. Esse padrão confirma que a calibração contribui para uma redistribuição mais realista dos escores de confiança, permitindo explorar diferentes regimes operacionais do modelo.

Adicionalmente, verifica-se que a variação dos limiares impacta não apenas as métricas de desempenho, mas também o volume de dados selecionados para o treinamento, influenciando diretamente a capacidade de generalização do modelo. Iterações com maior quantidade de relatos não necessariamente resultam em melhor desempenho, como observado na segunda iteração, sugerindo a introdução de ruído proveniente de predições menos confiáveis. Por outro lado, limiares muito restritivos reduzem substancialmente o conjunto de dados, o que pode comprometer a robustez do modelo. Assim, os resultados indicam que, na ausência de metadados, a calibração por *Isotonic Regression* torna-se um mecanismo relevante para o controle entre qualidade e quantidade dos dados pseudo-rotulados, embora a definição adequada do limiar de confiança permaneça um fator crítico para o desempenho final.

Tabela V.15: Resultados das iterações de pseudo-rotulação realizadas sem a utilização de metadados, mas com aplicação da técnica de calibração *Isotonic Regression*. A tabela apresenta os valores das iterações com diferentes limiares de confiança, incluindo a quantidade de relatos mantidos e descartados e as métricas de desempenho (F1, precisão e revocação).

Iteração	Limiar	Mantidos	Descartados	F1	Precisão	Revocação
1	0,7	151.354	215.500	63,0%	68,4%	58,4%
2	0,8	209.817	5.683	57,2%	67,7%	49,6%
3	0,3	908	4.775	58,0%	67,6%	50,8%
4	0,2	629	4.146	59,7%	58,0%	61,5%
5	0,2	2.630	1.516	58,8%	56,7%	61,1%

A Figura V.10 apresenta a matriz de confusão obtida na avaliação do modelo sem utilização de metadados, mas com aplicação da técnica de calibração *Isotonic Regression*. Observa-se que a classe *Localização* apresentou o maior número de classificações corretas, com 827 ocorrências corretamente identificadas. A classe **Organização** apresentou 114 classificações corretas, com baixo número de confusões com as demais categorias.

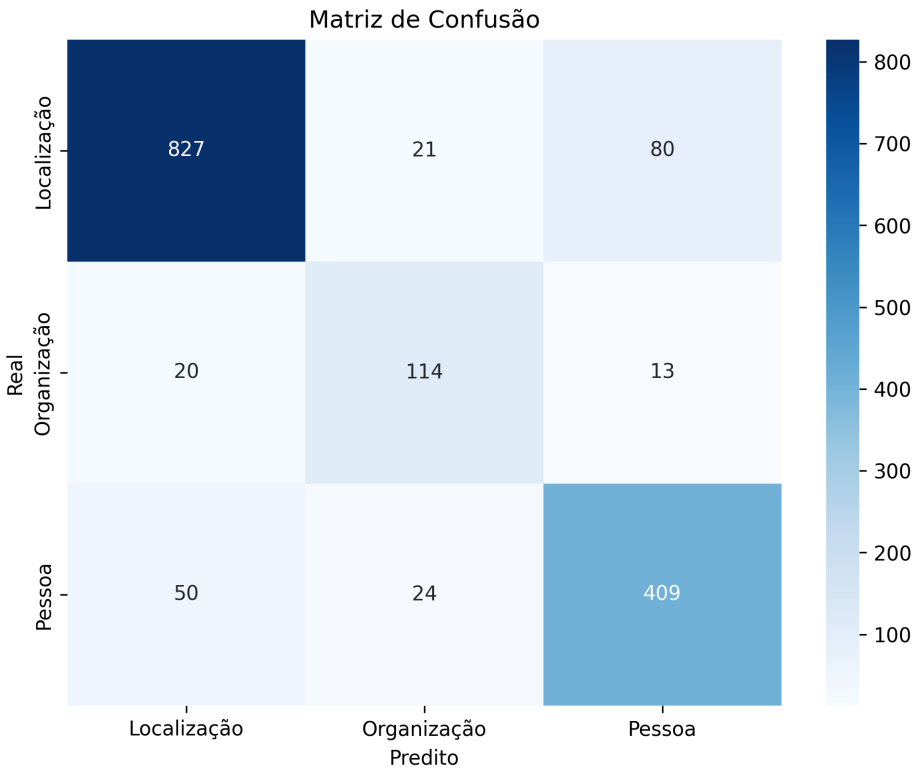


Figura V.10: Matriz de confusão obtida na avaliação do modelo treinado sem utilização de metadados, mas com aplicação da técnica de calibração *Isotonic Regression* na primeira iteração disponível. A figura apresenta a distribuição das predições corretas e incorretas entre as classes **Localização**, **Organização** e **Pessoa**.

V.8 Discussão

O modelo GLiNER apresentou o melhor desempenho no domínio dos relatos criminais do Disque Denúncia e foi avaliado em duas versões: *multilanguage* e *large*. A versão *multilanguage* é treinada em um *corpus* permitindo sua aplicação em diferentes idiomas e utiliza aproximadamente 209 milhões de parâmetros pré-treinados. A versão *large*, por sua vez, é treinada primariamente em inglês e possui cerca de 459 milhões de parâmetros.

Em comparação com os trabalhos relacionados apresentados na Seção III.3.1, os resultados obtidos nesta pesquisa buscam aprofundar o entendimento e a avaliação de metáforas, eufemismos e expressões informais presentes nos relatos criminais, conforme discutido em Zhu and Bhat [2021] e Rothe et al. [2023]. Nesse contexto, a Tabela V.8 apresenta uma síntese comparativa do desempenho dos modelos avaliados, incluindo abordagens previamente aplicadas ao mesmo domínio, como o modelo LSTM, que obteve F1 de 0,28 no estudo descrito em Figueiredo et al. [2023].

Observa-se que os modelos baseados no GLiNER alcançam os melhores resultados globais, com destaque para a versão *multilanguage* (F1 de 0,60), seguidos pelo Bertimbau (F1 de 0,56). Em contraste, modelos generalistas, como GPT e Gemini, mesmo quando submetidos ao ajuste fino, apresentam desempenho inferior, especialmente em termos de revocação, evidenciando limitações

na captura de padrões linguísticos específicos do domínio. Esses resultados reforçam a importância do uso de modelos especializados ou adaptados ao contexto da tarefa, particularmente em cenários caracterizados por elevada informalidade e variabilidade linguística.

Ao transpor essas abordagens para o contexto da língua portuguesa, este trabalho amplia a análise desses fenômenos linguísticos em um domínio específico e sensível, marcado por variações sociolinguísticas e uso recorrente de gírias. Observou-se que modelos baseados na arquitetura BERT apresentaram ganhos mais expressivos, especialmente na capacidade de capturar nuances contextuais e padrões implícitos de linguagem, resultado que se alinha às evidências apresentadas em Caselli et al. [2021].

Dessa forma, os avanços verificados nesta dissertação permitem estabelecer uma comparação direta com estudos desenvolvidos na língua inglesa, demonstrando que o uso de modelos baseados em *Transformers* também proporciona ganhos significativos no português, ainda que adaptados às particularidades linguísticas e discursivas dos relatos criminais analisados.

O desempenho superior do GLiNER em comparação aos modelos LSTM, Gemini e GPT, demonstrado na Tabela V.8, pode ser explicado pelo fato de que estes são modelos generalistas, sem foco específico em tarefas de REN. Isso reduz sua precisão em contextos especializados, como o dos relatos criminais. O Spacy, utilizando o modelo *pt-core-news-lg*, apresentou desempenho superior aos modelos generalistas por possuir componentes voltados diretamente ao reconhecimento de entidades. O modelo mais próximo ao desempenho do GLiNER foi o BERTimbau, que compartilha a arquitetura baseada em BERT; porém, não alcançou o mesmo nível de eficácia, indicando menor adaptação ao *corpus* testado.

A vantagem do GLiNER também está associada à sua arquitetura mais leve e atualizada em relação ao BERT-base. O BERTimbau utilizado neste estudo (*bert-base-portuguese-cased*) possui aproximadamente 110 milhões de parâmetros, quantidade inferior à das variantes do GLiNER empregadas, o que contribui para a diferença observada nos resultados.

A pseudo-rotulação, representada pelas Tabelas V.10, V.11, V.12, V.13, V.14 e V.15, apresentou ganhos modestos em relação aos resultados obtidos sem essa técnica que está disponível na Tabela V.8, indicando que, em alguns cenários, não é necessário um grande número de relatos para alcançar melhorias significativas. Contudo, devido ao volume elevado de dados no *corpus* DD_{large}, o custo computacional do processo foi alto, o que reduziu sua relação custo-benefício.

De forma complementar, os resultados de cada experimento indicam variações moderadas entre os cenários avaliados. No experimento com uso de metadados e sem calibração, o melhor resultado foi obtido na quarta iteração, com *F1* de 67,0%. Quando os metadados foram combinados com calibração por *Temperature Scaling*, o melhor desempenho ocorreu na primeira iteração, com *F1* de 65,3%, utilizando um volume menor de relatos. Já com *Isotonic Regression*, o melhor resultado

também atingiu *F1* de 67,0%, ao considerar todos os relatos disponíveis na última iteração.

Nos experimentos sem utilização de metadados, observou-se desempenho ligeiramente inferior. Sem calibração, o melhor resultado foi de 65,4% de *F1*, obtido na segunda iteração. Com calibração por *Temperature Scaling*, o maior *F1* foi de 62,4%, enquanto a calibração por *Isotonic Regression* alcançou 63,0%. Esses resultados sugerem que os metadados contribuíram de forma pontual para o desempenho, enquanto a calibração teve maior impacto na seleção dos relatos e na redução do volume necessário para o treinamento.

Neste trabalho, a confiança do relato, citada na Seção IV.3.5, foi definida como a média das confianças atribuídas às entidades reconhecidas no texto durante os ciclos de treinamento autoiterativo. Essa abordagem apresenta a vantagem de ser simples, interpretável e coerente com a estrutura das predições do modelo, permitindo ordenar os relatos de acordo com o grau médio de confiança das entidades identificadas. No entanto, essa estratégia possui limitações, uma vez que a agregação por média pode mascarar variações internas entre entidades, especialmente em relatos com poucas ocorrências, o que pode levar ao descarte de exemplos potencialmente relevantes e, consequentemente, impactar negativamente o processo de generalização e o enriquecimento do treinamento autoiterativo.

Os melhores resultados, descritos nas Tabelas V.10, V.11 e V.12, foram observados quando mais de 298.614 relatos foram utilizados juntamente com metadados, sem a aplicação de técnicas de calibração. Entretanto, a calibração permitiu alcançar desempenhos semelhantes utilizando apenas 34.244 relatos. Isso demonstra que a calibração desempenha papel fundamental na obtenção de modelos eficientes, ao mesmo tempo reduzindo substancialmente o custo computacional.

Assim, a pseudo-rotulação (Tabelas V.11 e V.12) mostrou desempenho superior quando combinada com metadados e técnicas de calibração em comparação com a Tabela V.8. Nos cenários sem essas estratégias, observou-se desempenho similar, porém com um número muito maior de relatos mantidos nos treinamentos, implicando maior tempo de ajuste fino e menor eficiência do processo.

Entretanto, essa estratégia também apresenta algumas limitações. Relatos muito curtos, contendo poucas entidades, podem apresentar valores médios de confiança elevados mesmo quando a evidência contextual é limitada. Por outro lado, relatos mais longos, que incluem maior diversidade de entidades e descrições, podem apresentar médias menores devido à presença de algumas entidades com menor confiança, ainda que o conjunto da informação seja relevante. Dessa forma, embora a média das confianças seja uma métrica prática para seleção de relatos, futuras investigações poderiam explorar alternativas complementares, como métricas baseadas na quantidade de entidades reconhecidas, na variância dos escores de confiança ou em combinações ponderadas dessas medidas.

Foram adotados diferentes limiares de confiança do relato, permitindo controlar o nível de confiabilidade dos dados incorporados ao processo de treinamento autoiterativo. Essa estratégia

visa equilibrar dois fatores centrais: a qualidade dos pseudo-rótulos e o volume de dados utilizados no ajuste fino. Conforme evidenciado nas Tabelas V.10, V.11, V.12, V.13, V.14 e V.15, limiares mais elevados tendem a selecionar relatos com maior confiabilidade, reduzindo a probabilidade de ruído no treinamento.

Entretanto, essa escolha também implica uma redução significativa no número de exemplos disponíveis, o que pode limitar a diversidade das entidades aprendidas pelo modelo. Em contrapartida, limiares mais baixos ampliam o conjunto de dados utilizados, mas aumentam o risco de incorporação de predições incorretas. De forma geral, os resultados observados ao longo das iterações indicam a existência de um ponto de equilíbrio entre esses fatores, no qual limiares intermediários permitem manter uma quantidade adequada de relatos, ao mesmo tempo em que preservam um nível satisfatório de qualidade nas anotações geradas automaticamente.

Embora as métricas tenham variado pouco entre iterações e cenários, a quantidade de relatos selecionados variou significativamente entre os diferentes limiares, influenciando diretamente o custo computacional. Nesse contexto, a calibração desempenha um papel fundamental ao tornar o processo de seleção mais criterioso e eficiente, contribuindo para a redução do viés de superconfiança do modelo e promovendo uma redistribuição mais adequada dos escores de confiança. Como resultado, observa-se uma melhoria na qualidade da seleção dos relatos criminais utilizados no processo de pseudo-rotulação, favorecendo a incorporação de exemplos mais representativos e confiáveis para o treinamento autoiterativo.

Os resultados da utilização dos metadados presentes nas Tabelas V.10, V.11 e V.12, embora tenham oferecido algum ganho, mostraram impacto limitado, pois frequentemente apenas confirmam informações já presentes nos relatos e reconhecidas pelo modelo que estava com superconfiança. Assim, seu efeito foi menos perceptível no desempenho final do GLiNER.

A calibração, por outro lado, teve impacto relevante ao reduzir o viés de superconfiança e melhorar o alinhamento entre confiança e acurácia, mesmo com poucos relatos criminais disponíveis para fazer a análise. O *Temperature Scaling* obteve os melhores resultados, aproximando-se mais da curva de calibração ideal, enquanto a *Isotonic Regression* apresentou maior variação nos valores de confiança.

Modelos mais bem calibrados descritos na Tabela V.11 demonstraram maior eficiência no reconhecimento de entidades no contexto da segurança pública, proporcionando maior diversificação e confiabilidade nas entidades extraídas dos relatos. A aplicação de *Temperature Scaling* contribuiu para alinhar os scores de confiança à acurácia real, reduzindo discrepâncias entre métricas como precisão e F1.

Além disso, observou-se melhor equilíbrio entre precision e recall, com redução de falsos positivos e negativos. Os resultados também indicam maior robustez entre diferentes tipos de entidades, su-

gerindo melhor generalização em cenários com dados desbalanceados. Os modelos calibrados apresentam desempenho mais consistente e maior confiabilidade para aplicações em contextos críticos.

Algumas limitações foram identificadas, como o tamanho reduzido do conjunto de relatos utilizados para calibração (100 relatos criminais) e a heterogeneidade das entidades presentes nos *corpora*, com predominância de entidades de Localização em detrimento das demais. Essas restrições impactaram negativamente o desempenho e a capacidade de diversificação da calibração dos modelos.

A alta predominância de entidades de Localização direcionou o processo de treinamento,tanto no ajuste fino quanto na pseudo-rotulação, para um melhor desempenho nessa classe específica, em prejuízo de outras, como Organização e Pessoa. Como consequência, os modelos apresentaram menor capacidade de generalização, com desempenho menos consistente entre diferentes tipos de entidades e um menor aprendizado e diversificação nos reconhecimentos.

Capítulo VI Conclusão

Neste capítulo apresentamos a conclusão do trabalho e discutimos possíveis trabalhos futuros. Na Seção VI.1 procedemos uma breve análise retrospectiva e analisamos os resultados alcançados. Na Seção VI.2 apresentamos algumas sugestões de possíveis trabalhos futuros.

Todos os experimentos descritos neste trabalho foram desenvolvidos com a linguagem de programação Python 3.12. Para fins de reprodutibilidade, nosso código-fonte está disponível publicamente em https://github.com/ailab-cefet-rj/nerdd. Os procedimentos de instalação e de replicação dos experimentos estão descritos no próprio repositório.

VI.1 Análise Retrospectiva

Conforme apresentado na Seção I.3, este trabalho propôs um método de auto-treinamento iterativo para o reconhecimento de entidades nomeadas em relatos criminais. Os resultados experimentais confirmaram a eficácia da pseudo-rotulação como estratégia para ampliar os dados de treinamento, especialmente quando combinada ao uso de metadados e técnicas de calibração. Além disso, a construção e validação dos corpora DD_{small} e DD_{large} possibilitaram o estabelecimento de um benchmark para a avaliação comparativa dos modelos.

O método proposto foi desenvolvido a partir do GLiNER e demonstrou que a incorporação de metadados contribui para aumentar a qualidade das predições e a capacidade de generalização do modelo. Os experimentos indicaram que a pseudo-rotulação produziu ganhos superiores aos cenários sem sua utilização, reduzindo a dependência de validação manual e permitindo o aproveitamento de grandes volumes de relatos criminais.

Os experimentos também evidenciaram um comportamento de superconfiança nas predições, tornando necessária a aplicação de técnicas de calibração. Entre os métodos avaliados, o *Temperature Scaling* apresentou os melhores resultados, promovendo melhor alinhamento entre confiança e acurácia e tornando o processo de pseudo-rotulação mais confiável.

O melhor resultado foi obtido na quarta iteração sem calibração, alcançando 67% de F1 após o processamento de 362.808 relatos. Entretanto, a primeira iteração com Temperature Scaling atingiu 65,3% de F1 utilizando apenas 34.244 relatos, evidenciando uma relação mais favorável entre desempenho e custo computacional.

A comparação com os modelos de referência demonstrou que abordagens especializadas e submetidas a processos de ajuste fino tendem a apresentar desempenho superior no REN. Os resultados observados indicam que a adaptação dos modelos ao domínio específico dos relatos criminais é um fator determinante para a obtenção de melhores métricas de desempenho, sobretudo em cenários caracterizados pela presença de linguagem informal, abreviações e variações linguísticas. Dessa forma, os experimentos reforçam a importância da utilização de estratégias de especialização e treinamento direcionado para aplicações de reconhecimento de entidades em domínios específicos.

Como principais contribuições deste trabalho destacam-se a proposição de um método de autotreinamento iterativo baseado em pseudo-rotulação para o reconhecimento de entidades em relatos criminais, a incorporação de metadados ao processo de modelagem, a aplicação de técnicas de calibração para reduzir a superconfiança das predições e a análise da relação entre desempenho e custo computacional ao longo das iterações de treinamento. Além disso, o estudo contribui com a construção e validação de recursos experimentais que possibilitaram a avaliação sistemática dos modelos em um domínio especializado.

Como limitações, destacam-se o número reduzido de exemplos utilizados na calibração, a dependência parcial de validação manual e o elevado custo computacional associado às etapas de expansão do conjunto de treinamento. Apesar dessas restrições, a metodologia proposta não se limita ao contexto dos relatos criminais, podendo ser adaptada e avaliada em diferentes aplicações de REN e em outros tipos de dados textuais.

VI.2 Trabalhos Futuros

Como trabalhos futuros, pode-se explorar diferentes estratégias de integração dos metadados ao processo de ajuste fino, incorporando-os de forma sistemática ao pipeline de diferentes modelos do tipo LLM. Outra direção relevante consiste em aprimorar a utilização da pseudo-rotulação, buscando maior eficiência e qualidade nas iterações do processo autoiterativo, por meio de estratégias que permitam ativar e aproveitar de forma mais ampla os dados disponíveis nos *corpora*. Além disso, a utilização de um volume maior de relatos criminais validados manualmente representa uma oportunidade importante para expandir o processo de calibração dos modelos e, consequentemente, fortalecer a robustez e a confiabilidade dos resultados obtidos na validação experimental.

Para ampliar os estudos, conforme destacado na Figura IV.3, pode-se adotar uma estratégia de particionamento dos dados na proporção de 70% para treinamento e 30% para validação em cada um dos *folds* gerados. Essa abordagem possibilita uma verificação mais robusta do desempenho do modelo, além de promover maior diversidade nas partições utilizadas ao longo dos experimentos, contribuindo para uma avaliação mais consistente e confiável dos resultados produzidos.

Outra direção promissora consiste no desenvolvimento de um dicionário estruturado de enti-

dades e de uma aplicação dedicada, na qual novos textos provenientes de relatos criminais possam ser automaticamente processados pelos modelos desenvolvidos neste trabalho. Essa aplicação permitiria disponibilizar, de forma automatizada, as entidades identificadas com elevado nível de assertividade, contribuindo tanto para a padronização das análises quanto para o aprimoramento contínuo do sistema proposto.

Adicionalmente, recomenda-se a aplicação dos modelos em análises segmentadas por categorias, considerando separadamente os diferentes tipos de relatos criminais. Essa abordagem permitiria avaliar de forma mais precisa o comportamento dos modelos em contextos específicos. Torna-se possível também integrar bases externas ao processo de análise, incorporando informações adicionais, como dados geográficos, registros de organizações e outras entidades relevantes presentes em diferentes contextos da sociedade.

A utilização dessas bases pode contribuir para ampliar o conhecimento extraído dos relatos criminais, permitindo identificar relações entre locais, pessoas e organizações mencionadas. Dessa forma, torna-se possível estruturar os dados presentes nos relatos de maneira mais consistente e assertiva, possibilitando que essas informações apoiem análises e aplicações voltadas ao suporte das atividades das forças de segurança pública.

A publicação deste trabalho já tem motivado o desenvolvimento de iniciativas voltadas à aplicação prática dos modelos propostos, especialmente na análise e visualização de informações extraídas dos relatos do Disque Denúncia. Entre essas iniciativas, destaca-se o projeto *Crime Heatmap*, desenvolvido pelo aluno **Daniel Salvador** e disponível em https://github.com/DanielMT7/crime-heatmap, que utiliza entidades reconhecidas automaticamente para apoiar a análise espacial de ocorrências criminais. Além disso, outros alunos e pesquisadores colaboradores têm participado de projetos derivados desta pesquisa, explorando diferentes formas de utilização dos modelos desenvolvidos, o que evidencia o potencial de expansão e aplicação da metodologia proposta em novos contextos de análise de dados criminais.

Referências Bibliográficas

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*. 15

Arazo, E., Ortego, D., Albert, P., O'Connor, N., and McGuinness, K. (2020). Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)*, pages 908–917. 21

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171. 8

Caselli, T., Basile, V., Mitrović, J., and Granitzer, M. (2021). Hatebert: Retraining bert for abusive language detection in english. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25. 34, 36, 89

Centro de Estudos de Segurança e Cidadania (2025). Relatório de segurança pública do estado do rio de janeiro. Acessado em: 11/01/2026. 2

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537. 8

da Silva, D. and Melo, G. (2020). Estudos para classificação de denúncias por canais digitais para o disque denúncia. Trabalho de Conclusão de Curso. Universidade do Estado do Rio de Janeiro (UERJ). 3, 40

da Silva, F. S., Nogueira, R., and Lotufo, R. (2023). Bertimbau: Pretrained bert models for brazilian portuguese. *Language Resources and Evaluation*, 57(1):209–231. 31, 36, 70

Devlin, J. et al. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*. 7, 14, 15, 19

Dharviyanti, N. A. D. and Wilantika, N. (2024). Rule-based named entity recognition for crime information extraction through online news sites. In *Proceedings of the 2024 International Conference on Information Technology Research and Innovation (ICITRI)*, pages 99–104. IEEE. 33, 36

Disque Denúncia (2023). Disque denúncia. Acessado em: 10/05/2025. 2

Elangovan, A., He, J., and Verspoor, K. (2021). Memorization vs. generalization: Quantifying data leakage in nlp performance evaluation. *arXiv preprint arXiv:2102.01818*. Acessado em: 08/08/2025. 21

Ferreira, F., Duarte, J., and Ugulino, W. (2022). Automated statistics extraction of public security events reported through microtexts on social networks. In *Proceedings of the XVIII Brazilian Symposium on Information Systems*, pages 1–7. 32, 36

Figueiredo, K., Tomé, Y., Barreto, L., and Brito, W. A. T. (2023). Estudo de modelo deep learning para reconhecimento de entidades nomeadas na segurança pública. Trabalho acadêmico. 3, 69, 88

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378. 23

Gomes Jr., J., Bernardino, H. S., de Souza, J. F., and Rajabi, E. (2023). Indexing, enriching, and understanding brazilian missing person cases from data of distributed repositories on the web. *AI & SOCIETY*, 38(2):565–579. 32, 36, 77

Gonzalez Hernandez, F., Nguyen, Q., Smith, V. C., Cordero, J. A., Ballester, M. R., Duran, M., Solé, A., Chotsiri, P., Wattanakul, T., Mundin, G., et al. (2024). Named entity recognition of pharmacokinetic parameters in the scientific literature. *Scientific Reports*, 14(1):23485. 20

Goutte, C. and Gaussier, E. (2005). A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *Proceedings of the European Conference on Information Retrieval (ECIR 2005)*, pages 345–359. Springer. 10

Grishman, R. and Sundheim, B. M. (1996). Message understanding conference-6: A brief history. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING 1996), Volume 1*. 8

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, pages 1321–1330. 23, 56

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. *Proceedings of the 56th Annual Meeting of the ACL*. 19

Instituto de Pesquisa Econômica Aplicada (2024). Atlas da violência 2024. Acessado em: 12/06/2025. 1

Jeong, Y. and Kim, E. (2022). Scideberta: Learning deberta for science and technology documents and fine-tuning information extraction tasks. *IEEE Access*, 10:60805–60813. 8

Keraghel, I., Morbieu, S., and Nadif, M. (2024). A survey on recent advances in named entity recognition. *arXiv preprint arXiv:2401.10825*. Acessado em: 08/05/2025. 16

Labusch, K., Neudecker, C., and Zellhöfer, D. (2019). Bert for named entity recognition in contemporary and historical german. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS)*, pages 9–11. 8

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, Williamstown, MA. 8

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*. Acessado em: 10/07/2025. 9

Lee, D.-H. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)*. 21

Li, J. et al. (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*. 8, 10, 11

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600. 23

Naeini, M. P. et al. (2015). Obtaining well calibrated probabilities. *Proceedings of the AAAI Conference on Artificial Intelligence*. 24

Nguyen, L. M. (2024). GLiNER: A generalist model for named entity recognition using bidirectional transformers. Acessado em: 08/03/2025. 16

Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML 2005)*, pages 625–632. 23, 24

Otal, H. T., Stern, E., and Canbaz, M. A. (2024). Llm-assisted crisis management: Building advanced llm platforms for effective emergency response and public collaboration. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 851–859. IEEE. 34, 36

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., et al. (2021). The prisma 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372:n71. 29

Pahuja, V., Yadav, N., and Saini, R. (2021). A systematic review of named entity recognition in social media. *arXiv preprint arXiv:2109.08961*. Acessado em: 08/09/2025. 21, 22

Perez, G. et al. (2024). The costs of crime and violence in latin america. Acessado em: 12/08/2025. 1

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Technical Report MSR-TR-99-37, Microsoft Research. 23

Powers, D. (2020). Evaluation: From precision, recall and f-measure to roc. *Journal of Machine Learning Technologies*. 10

Ratinov, L. and Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009)*, pages 147–155. 8, 9

Rothe, M., Lath, R., Kumar, D., Yadav, P., and Aylani, A. (2023). Slang language detection and identification in text. In *Proceedings of the 14th International Conference on Computing Communication and Networking Technologies (ICCCNT 2023)*, pages 1–5. IEEE. 33, 36, 88

Shi, Y. and Kimura, M. (2024). Bert-based models with attention mechanism and lambda layer for biomedical named entity recognition. In *Proceedings of the 16th International Conference on Machine Learning and Computing*, pages 536–544. 8

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures. *Information Processing & Management*. 10

Souza, F. et al. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. *Brazilian Symposium on Artificial Intelligence*. 7, 69

Stepanov, D., Zaratiana, H., and Sennrich, R. (2024). GLiNER-multitask: Multitask learning for generalist named entity recognition. *arXiv preprint arXiv:2404.04593*. Acessado em: 08/03/2025. 16, 18

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Joulin, A., Grave, E., and Lample, G. (2023). LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*. Acessado em: 20/08/2025. 14

Tsang, e. a. (2022). A review of pseudo-labeling. *IEEE Access*. 22

Vacareanu, R., Noriega-Atala, E., Hahn-Powell, G., Valenzuela-Escárcega, M. A., and Surdeanu, M. (2024). Active learning design choices for named entity recognition with transformers. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 321–334. 8

Vaswani, A. et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*. 13, 14

Wang, X., Kordjamshidi, P., and Roth, D. (2019). Crossweigh: Training named entity tagger from imperfect annotations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 5153–5163. 21

Yadav, V. and Bethard, S. (2019). A survey on recent advances in named entity recognition. *Artificial Intelligence Review*. 8, 9, 11

Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699. 23

Zaratiana, e. a. (2023). GLiNER: Generalist named entity recognition. *arXiv preprint arXiv:2311.08526*. Acessado em: 20/09/2025. 15, 16, 17, 18, 19

Zhu, W. and Bhat, S. (2021). Euphemistic phrase detection by masked language model. *arXiv preprint arXiv:2109.04666*. Acessado em: 08/09/2025. 32, 36, 88

Zhu, X. J. (2005). Semi-supervised learning literature survey. Technical Report TR 1530, University of Wisconsin–Madison, Department of Computer Sciences. 21

Zilliz (2024). GLiNER: Generalist model for named entity recognition using bidirectional transformer. Acessado em: 08/03/2025. 16, 18

Apêndice A Nuvens de Palavras



Figura A.1: Nuvem de palavras gerada a partir das entidades extraídas dos relatos criminais no corpus $DD_{\textbf{small}}$. As entidades mais frequentes são representadas com maior destaque por meio do aumento proporcional do tamanho da fonte.



Figura A.2: Nuvem de palavras gerada a partir das entidades de Localização extraídas dos relatos criminais no corpus $DD_{\textbf{small}}$. As entidades mais frequentes são representadas com maior destaque por meio do aumento proporcional do tamanho da fonte.



Figura A.3: Nuvem de palavras gerada a partir das entidades de Pessoa extraídas dos relatos criminais no corpus $DD_{\textbf{small}}$. As entidades mais frequentes são representadas com maior destaque por meio do aumento proporcional do tamanho da fonte.



Figura A.4: Nuvem de palavras gerada a partir das entidades de Pessoa extraídas dos relatos criminais no corpus $DD_{\textbf{small}}$. As entidades mais frequentes são representadas com maior destaque por meio do aumento proporcional do tamanho da fonte.



Figura A.5: Nuvem de palavras gerada a partir das entidades extraídas dos relatos criminais no corpus $DD_{\textbf{large}}$. As entidades mais frequentes são representadas com maior destaque por meio do aumento proporcional do tamanho da fonte.



Figura A.6: Nuvem de palavras gerada a partir das entidades de Localização extraídas dos relatos criminais no corpus $DD_{\mathbf{large}}$. As entidades mais frequentes são representadas com maior destaque por meio do aumento proporcional do tamanho da fonte.



Figura A.7: Nuvem de palavras gerada a partir das entidades de Pessoa extraídas dos relatos criminais no corpus $DD_{\mathbf{large}}$. As entidades mais frequentes são representadas com maior destaque por meio do aumento proporcional do tamanho da fonte.



Figura A.8: Nuvem de palavras gerada a partir das entidades de Organização extraídas dos relatos criminais no corpus $DD_{\textbf{large}}$. As entidades mais frequentes são representadas com maior destaque por meio do aumento proporcional do tamanho da fonte.