ARTHUR SILVEIRA FRANCO

AVALIAÇÃO EXPERIMENTAL DE BASES DE DADOS PARA RECONHECIMENTO DE ENTIDADES NOMEADAS NA LÍNGUA PORTUGUESA

LAVRAS – MG
2023

ARTHUR SILVEIRA FRANCO

AVALIAÇÃO EXPERIMENTAL DE BASES DE DADOS PARA RECONHECIMENTO DE ENTIDADES NOMEADAS NA LÍNGUA PORTUGUESA

TCC apresentado à Universidade Federal de Lavras, como parte das exigências do Trabalho de Conclusão do Curso de Bacharelado em Ciência da Computação, área de concentração em Processamento de Linguagem Natural, para a obtenção do título de Bacharel.

Prof. Denilson Alves Pereira, Ph.D.
Orientador

LAVRAS – MG
2023

Ficha Catalográfica preparada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Franco, Arthur Silveira
Avaliação Experimental de Bases de Dados para Reconhecimento de Entidades Nomeadas na Língua Portuguesa / Arthur Silveira Franco. 1^a ed. rev., atual. e ampl. – Lavras : UFLA, 2023.
27 p. : il.

TCC(graduação)–Universidade Federal de Lavras, 2023.
Orientador: Prof. Denilson Alves Pereira, Ph.D..
Bibliografia.

1. TCC. 2. Monografia. 3. Dissertação. 4. Tese. 5. Trabalho Científico – Normas. I. Universidade Federal de Lavras. II. Título.

CDD-808.066

ARTHUR SILVEIRA FRANCO

AVALIAÇÃO EXPERIMENTAL DE BASES DE DADOS PARA RECONHECIMENTO DE ENTIDADES NOMEADAS NA LÍNGUA PORTUGUESA

TCC apresentado à Universidade Federal de Lavras, como parte das exigências do Trabalho de Conclusão do Curso de Bacharelado em Ciência da Computação, área de concentração em Processamento de Linguagem Natural, para a obtenção do título de Bacharel.

APROVADA em 28 de julho de 2023.

Prof. Dra. Marluce Rodrigues Pereira	UFLA
Prof. Dra. Paula Christina Figueira Cardoso	UFLA

gov.br
Documento assinado digitalmente
DENILSON ALVES PEREIRA
Data: 03/08/2023 10:02:46-0300
Verifique em https://validar.iti.gov.br

Prof. Denilson Alves Pereira, Ph.D.
Orientador

LAVRAS – MG
2023

RESUMO

O Reconhecimento de Entidades Nomeadas é uma tarefa de extração de informação extremamente importante, exercendo papel chave em diversas áreas do Processamento de Linguagem Natural, como na mineração de opinião, perguntas e respostas e tradução automática. Embora tenham sido alcançados avanços significativos nessa área, alguns idiomas, incluindo o Português, ainda enfrentam escassez de recursos linguísticos, como bases de dados rotuladas manualmente. Além disso, a falta de um padrão de partições predefinidas dificulta a replicabilidade de experimentos e a comparação justa entre diferentes abordagens de Reconhecimento de Entidades Nomeadas. Este trabalho aborda essa lacuna e propõe uma metodologia de particionamento aplicada a sete bases de dados para Reconhecimento de Entidades Nomeadas em Português, que resultou em 10 partições disjuntas para cada uma das mesmas. Ademais, também é apresentado e discutido o desempenho de um classificador baseado no modelo de linguagem BERTimbau nas bases de dados utilizadas.

Palavras-chave: Reconhecimento de Entidades Nomeadas. Avaliação Experimental. Partição de Bases de Dados.

ABSTRACT

Named Entity Recognition is an extremely important information extraction task, playing a key role in various areas of Natural Language Processing, such as opinion mining, question answering, and machine translation. Despite significant advancements in this field, some languages, including Portuguese, still face a scarcity of linguistic resources, such as manually labeled datasets. Moreover, the lack of a predefined partitioning standard hinders the replicability of experiments and fair comparison between different Named Entity Recognition approaches. This work addresses this gap and proposes a partitioning methodology applied to seven datasets for Named Entity Recognition in Portuguese, resulting in 10 disjoint partitions for each of them. Additionally, the performance of a classifier based on the BERTimbau language model on the utilized datasets is presented and discussed.

Keywords: Named Entity Recognition. Experimental Evaluation. Dataset Partitioning .

Sumário

1 Introdução **2**

2 Trabalhos Relacionados **4**

3 Metodologia de Preparação das Bases de Dados **6**

3.1 Pré-processamento 7

3.2 Datasets 8

3.2.1 HAREM 8

3.2.2 Paramopama 11

3.2.3 LeNER-Br 12

3.2.4 UlyssesNER-Br 13

3.2.5 CachacaNER 15

4 Avaliação Experimental **16**

4.1 Configuração Experimental 16

4.2 Métricas para Avaliação 19

4.3 Resultados e Discussão 19

5 Conclusão e Trabalhos Futuros **25**

Referências **25**

Avaliação Experimental de Bases de Dados para Reconhecimento de Entidades Nomeadas na Língua Portuguesa

Arthur Silveira Franco[1], **Denilson Alves Pereira**[1]

[1]Departamento de Ciência da Computação

Universidade Federal de Lavras – Caixa postal 3037, Lavras, 37.200-900, MG, Brasil.

arthur.franco@estudante.ufla.br, denilsonpereira@ufla.br

Abstract. *Named Entity Recognition is an extremely important information extraction task, playing a key role in various areas of Natural Language Processing, such as opinion mining, question answering, and machine translation. Despite significant advancements in this field, some languages, including Portuguese, still face a scarcity of linguistic resources, such as manually labeled datasets. Moreover, the lack of a predefined partitioning standard hinders the replicability of experiments and fair comparison between different Named Entity Recognition approaches. This work addresses this gap and proposes a partitioning methodology applied to seven datasets for Named Entity Recognition in Portuguese, resulting in 10 disjoint partitions for each of them. Additionally, the performance of a classifier based on the BERTimbau language model on the utilized datasets is presented and discussed.*

Resumo. *O Reconhecimento de Entidades Nomeadas é uma tarefa de extração de informação extremamente importante, exercendo papel chave em diversas áreas do Processamento de Linguagem Natural, como na mineração de opinião, perguntas e respostas e tradução automática. Embora tenham sido alcançados avanços significativos nessa área, alguns idiomas, incluindo o Português, ainda enfrentam escassez de recursos linguísticos, como bases de dados rotuladas manualmente. Além disso, a falta de um padrão de partições predefinidas dificulta a replicabilidade de experimentos e a comparação justa entre diferentes abordagens de Reconhecimento de Entidades Nomeadas. Este trabalho aborda essa*

lacuna e propõe uma metodologia de particionamento aplicada a sete bases de dados para Reconhecimento de Entidades Nomeadas em Português, que resultou em 10 partições disjuntas para cada uma das mesmas. Ademais, também é apresentado e discutido o desempenho de um classificador baseado no modelo de linguagem BERTimbau nas bases de dados utilizadas.

1. Introdução

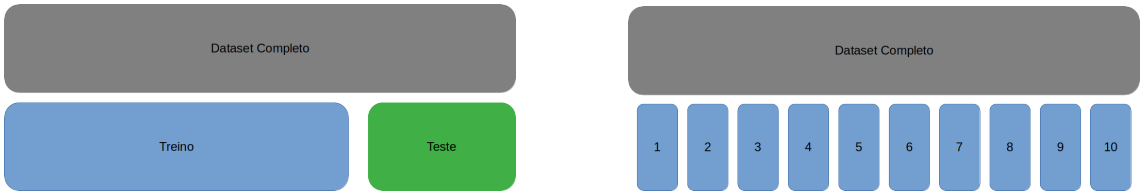
Reconhecimento de Entidades Nomeadas (NER, da sigla em inglês para *Named Entity Recognition*) é uma tarefa de extração de informação, onde elementos textuais são identificados e classificados do texto de entrada. Por exemplo, na frase: "Alan Turing, amplamente conhecido como o pai da ciência da computação, nasceu em Londres em 23 de junho de 1912", utilizando um sistema NER, é possível identificar as seguinte entidades nomeadas: "Alan Turing" (pessoa), "Londres" (localização) e "23 de junho de 1912" (data).

A tarefa é considerada multidisciplinar, exercendo papel chave em outras áreas do Processamento de Linguagem Natural (PLN) como na mineração de opinião, onde o NER pode ser aplicado para gerar análises mais detalhadas ao associar opiniões com entidades específicas, permitindo o entendimento de sentimento de consumidores sobre produtos e serviços [Marrero et al. 2013]. Também em tarefas de perguntas e respostas, o NER ajuda na extração de entidades relevantes nas entradas do usuário, que podem ser utilizadas para gerar respostas mais informativas e contextualizadas [Mollá et al. 2006]. A tarefa de tradução automática também pode ser aprimorada com o reconhecimento de entidades nomeadas como a identificação de nome de empresas, produtos e pessoas na língua fonte, para que o modelo de tradução possa traduzi-las de maneira consistente, mantendo o contexto e significado em todas as línguas [Babych and Hartley 2003].

O campo de pesquisa em PLN tem presenciado avanços significativos com a crescente disponibilidade de dados de treinamento e com a introdução de abordagens baseadas em Aprendizado de Máquina, sobretudo com os modelos de linguagem pré-treinados. Essas abordagens tiram proveito de grandes volumes de dados não rotulados para capturar padrões e informações semânticas valiosas que auxiliam em uma ampla gama de tarefas como o NER.

Apesar desses avanços, alguns idiomas, ainda contam com recursos linguísticos escassos, como base de dados rotuladas manualmente para tarefas específicas como NER. Para o Português, mesmo que em menos volume quando comparado ao Inglês, existem bases de dados disponíveis publicamente com textos de domínio geral ou específico, como o Jornalístico, Geológico, Clínico, dentre outros (Ver Tabela 3 em [Albuquerque et al. 2023]).

Um grande problema com as bases de dados disponíveis é a não existência de um padrão de partições disjuntas predefinidas (ilustradas na Figura 1), que permitam comparações mais justas entre diferentes abordagens com experimentos de validação cruzada ou o tradicional *hold-out*. Por exemplo, a base de dados mais utilizada para comparação de abordagens, HAREM [Santos et al. 2006] [Freitas et al. 2010], não possui nem mesmo um padrão de partições teste e treino, que é mais comum entre as bases de dados.



(a) Treino e Teste **(b) 10-Partes**

Figura 1. Partições Disjuntas

Com isso, cada trabalho convenciona um particionamento diferente. Por exemplo, [Souza et al. 2020] usa os conjuntos de dados Primeiro HAREM e Mini HAREM com treino e teste, respectivamente. Já [de Lima Santos. et al. 2021] combinam os três conjuntos e experimentam em diferentes cenários, variando de 70 a 95% dos dados para treino e 5 a 30% de dados de teste. Se essas partições não estão disponíveis publicamente, não é possível comparar a proposta de um novo modelo de NER com esses trabalhos, e acaba-se criando um novo particionamento.

Motivado por essa lacuna, este trabalho apresenta uma metodologia de particionamento em 10-partes disjuntas, aplicada a sete bases de dados para a tarefa de NER na língua portuguesa, bem como uma avaliação do desempenho de um classificador baseado no modelo de linguagem BERTimbau [Souza et al. 2020] para cada

uma das bases de dados. Os conjuntos de dados particionados estão disponíveis em `https://github.com/HarturFranco/NER-DS`.

O restante do artigo está estruturado da seguinte forma: a Seção 2 provê uma visão geral de trabalhos relacionados na literatura. A Seção 3 descreve em detalhes a metodologia de preparação e divisão das bases de dados selecionadas. A Seção 4 descreve o ambiente de execução e ferramentas utilizadas, a abordagem experimental e uma discussão sobre os resultados obtidos. Por fim, a Seção 5 conclui o artigo e destaca possíveis direções para trabalhos futuros.

2. Trabalhos Relacionados

Em [Yadav and Bethard 2018] foi apresentada uma pesquisa abrangente sobre arquiteturas de aprendizado profundo (*Deep Learning*) para o Reconhecimento de Entidades Nomeadas, as comparando com abordagens anteriores baseadas em *Feature Engenearing* e outros algoritmos de aprendizado supervisionados e semi-supervisionados. Os resultados mostraram que sistemas baseados em redes neurais superam as abordagens clássicas.

O modelo de Linguagem BERTimbau, utilizado no presente trabalho foi apresentado em [Souza et al. 2020]. O modelo é baseado na arquitetura do modelo BERT [Devlin et al. 2018] e pré-treinado em um grande volume de dados em Português. O artigo disponibiliza a avaliação do modelo para a tarefa de Reconhecimento de Entidades Nomeadas, na qual o mesmo supera os melhores resultados publicados anteriormente, melhorando a métrica F1 em 3,9 pontos no cenário total da base de dados HAREM [Santos et al. 2006]. O desempenho do modelo também foi analisado pelos autores em outras duas tarefas de PLN, sendo elas a similaridade entre sentenças (STS, da sigla em inglês para *Semantic Textual Similarity*) e o reconhecimento de implicação textual (RTE, da sigla em inglês para *Recognizing Textual Entailment*).

O treinamento de modelos de linguagem em idiomas específicos como o português demanda muito tempo e recursos computacionais extensivos. Com isso, os modelos de linguagem multilíngues surgem como uma opção mais viável, por meio de técnicas como o aprendizado por transferência e ajuste fino, onde o conhecimento adquirido pelo treinamento em diferentes idiomas é utilizado como ponto de partida e posteriormente refinado e adaptado para tarefas e idiomas específicos, como o Reconhecimento de Entidades No-

meadas na língua portuguesa . A pesquisa apresentada por [de Lima Santos. et al. 2021] explora o uso de modelos de linguagem multilíngues para o Reconhecimento de Entidades Nomeadas na língua portuguesa, comparando o desempenho dos mesmos ao estado-da-arte para modelos pré-treinados em Português. Para o aprimoramento e avaliação dos modelos, os autores construíram uma base de dados abrangente, contendo as três versões da base de dados HAREM. Os resultados apontam um desempenho superior dos modelos multilíngues após ajuste fino em grandes bases de dados.

Bases de dados rotuladas são de extrema importância para a criação e avaliação de abordagens de NER. Em [Albuquerque et al. 2023] foi realizado um mapeamento de técnicas, métodos e recursos para NER na língua portuguesa reportados nos últimos 12 anos. Foram incluídos 45 estudos primários na revisão, na qual 24 apresentaram novas bases de dados, ou versões atualizadas de bases de dados já existentes. Os autores apontam que o crescente número de estudos revelam maior interesse dos pesquisadores na área.

A criação de conjuntos de dados manualmente rotulados, conhecidos como coleções douradas (*golden-collections*), é altamente dispendiosa e trabalhosa. Ainda sim, estudos recentes apresentam um padrão de uso da estratégia de rotulação manual [Albuquerque et al. 2023].

No domínio da Geologia, [do Amaral et al. 2017] apresenta o processo de construção do GeoCorpus. O corpus foi produzido através da rotulação manual de 13 categorias de entidades geográficas, que englobam tempo geológico, rochas sedimentares, dentre outras categorias. A rotulação foi realizada em textos extraídos de teses, dissertações, artigos e boletins de Geociências da Petrobras no idioma português do Brasil.

Em [Consoli et al. 2020] foi utilizada uma versão revisada, denominada GeoCorpus-2. As modificações realizadas incluem a conversão de formato e a remoção de sentenças repetidas. O trabalho alcançou o estado da arte no reconhecimento de entidades nomeadas no domínio geológico, utilizando o modelo BiLSTM+CFR com utilização de *Word Embeddings* de domínio geral em combinação com *Flair Embedding Model* aprimorado para o domínio da geologia.

Uma alternativa à rotulação manual é a utilização de estratégias de rotulação au-

tomática, que, embora produzam corpora de qualidade inferior, denominadas coleções prata (*silver-collections*), desempenham um papel importante para tarefas como pré-treinamento de modelos de linguagem, especialmente pela rapidez e pelo potencial de cobertura de dados em larga escala. Nesse contexto, [Nothman et al. 2013] apresenta a criação do WikiNER, que explora os textos e estrutura do Wikipedia para rotular automaticamente uma grande quantidade de textos em diversos idiomas, incluindo o português. Para isso foi realizada a transformação do texto ancora de links entre artigos em entidades nomeadas atribuindo-as à categoria do artigo destino. As bases dados WikiANN [Pan et al. 2017] e SESAME [Menezes et al. 2019] também aproveitaram de textos e estruturas de enciclopédias online para rotulação automática de entidades nomeadas. Esses conjuntos, considerados de domínio geral, apresentam as categorias de local, pessoa e organização.

Das bases de dados utilizadas no presente trabalho, quatro são de domínio geral: Primeiro HAREM, Mini HAREM, Segundo HAREM [Santos et al. 2006] [Freitas et al. 2010] e Paramopama [Júnior et al. 2015]; duas no domínio Legal: LeNER-Br [Luz de Araujo et al. 2018] e UlyssesNER-Br [Albuquerque et al. 2022]; e uma no domínio da bebida brasileira Cachaça: CachacaNER [Silva et al. 2023]. Com exceção da base Paramopama, todas foram rotuladas manualmente. Essas bases de dados estão descritas em mais detalhes na Seção 3.2

3. Metodologia de Preparação das Bases de Dados

Nesta seção, descrevemos o processo de pré-processamento realizado nos datasets utilizados neste trabalho. O pré-processamento envolveu a conversão dos datasets para o formato IOB2, a quebra dos textos em sentenças e a aplicação de uma estratificação iterativa para criar partições disjuntas. A Figura 2 apresenta o fluxograma do processo.

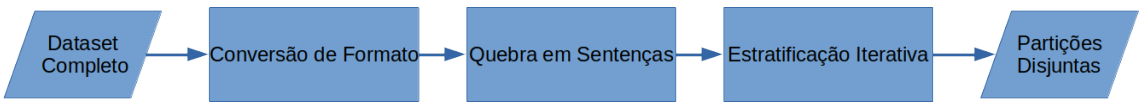


Figura 2. Fluxograma Pré-processamento

3.1. Pré-processamento

Inicialmente, os datasets estavam disponibilizados em diferentes formatos. Para garantir a padronização e facilitar a manipulação, realizamos a conversão de todos os datasets para o formato IOB2. Esse formato é derivado do esquema original IOB [Ramshaw and Marcus 1999]. Os formatos originais e os passos realizados para a conversão dos corpus são abordados individualmente na Subseção 3.2

```
Quem    O
é       O
Will    B-PESSOA
Carling I-PESSOA
?       O
```

Figura 3. Formato IOB2

A Figura 3 ilustra uma sentença no formato de texto utilizado, em que cada linha do arquivo é composta por um token pertencente à sentença e sua classe rotulada separada por uma tabulação. A figura mostra também o formato de anotação IOB2, no qual os rótulos "O" indicam que o token não pertence a nenhuma classe de entidades nomeadas, aqueles que começam por "B-" indicam o início de uma entidade nomeada e "I-" indica que o token faz parte da entidade nomeada iniciada com "B-".

Tendo o formato das bases padronizado, com intuito de facilitar a replicação de experimentos de validação cruzada, cada uma das bases foi dividida em 10 partições disjuntas. Para isso, foi implementado um algoritmo de estratificação iterativa baseado no algoritmo proposto por [Sechidis et al. 2011], visando manter a proporção das categorias de entidades nomeadas em todas partições.

A estratificação das bases de dados foi aplicada a nível de documento, com exceção da base de dados Paramopama [Júnior et al. 2015], que não possui divisão explícita por documento. Os documentos foram quebrados em sentenças, separadas por um sinal de pontuação de fim de sentença (.!?). A sequência de sentenças de um mesmo documento foi mantida em uma mesma partição, preservando o contexto do documento. Em contrapartida, a distribuição de entidades por partição é afetada na maioria das bases de dados, devido à variação de número de entidades por documento.

Como descrito no Pseudocódigo do algoritmo implementado (Figura 4), o mesmo leva em consideração a quantidade total de exemplos de cada categoria de entidade nomeada, atribuindo iterativamente um documento (ou sentença) que possui a categoria com menos exemplos na base completa à partição que menos possui exemplos da mesma.

```
Algorithm 1: Estratificação Iterativa
  Result: Partições Disjuntas
  Input  : initial_set
  while initial_set is not empty do
      min_label_class ← calculate_min_label_class(initial_set);
      examples_with_label ← select_examples_with_label(initial_set, min_label_class);
      foreach example in examples_with_label do
          selected_subset ← calculate_desired_subset( min_label_class);
          set_example_for_subset(example, selected_subset);
          initial_set ← remove_example(example, initial_set);
          recalculate_desire(selected_subset);
      end
      update_label_counts(initial_set);
  end
```

Figura 4. Pseudocódigo Estratificação Iterativa

3.2. Datasets

A seguir, são apresentadas as sete bases de dados utilizadas na avaliação, bem como etapas de pré-processamento específicas aplicadas a cada uma delas.

3.2.1. HAREM

A base dados HAREM é uma referência importante para o desenvolvimento e avaliação de sistemas de Reconhecimento de Entidades Nomeadas na língua portuguesa, sendo utilizado na academia como *benchmark* para comparações entre abordagens e demais bases de dados.

A base de dados leva o nome da campanha de avaliação conjunta para a área de Reconhecimento de Entidades Nomeadas, organizada pela Linguateca[1]. A Organização disponibiliza três coleções douradas resultantes dos eventos realizados.

As coleções estão disponíveis em formato XML, onde cada Entidade nomeada

[1]https://www.linguateca.pt/

contém atributos de categoria, tipo e subtipo. Além da quebra das categorias em tipos e subtipos, as bases de dados também possuem o recurso de anotações alternativas, mantendo a imprecisão do processo de anotação semântica, onde uma entidade nomeada pode pertencer a mais de um tipo e subtipos de categorias [Santos et al. 2006].

O Primeiro HAREM e Primeiro HAREM MINI (Mini HAREM) possuem o mesmo padrão de anotação. Já a segunda versão (Segundo HAREM) teve o padrão de anotação alterado, com melhorias e adições de novas características, como de atributos de correlação e tipo de correlação entre entidades nomeadas [Freitas et al. 2010].

Para o presente trabalho, durante o processo de pré-processamento, as coleções foram convertidas para o formato ilustrado na Subseção 3.1, com anotações no formato IOB2. Durante a conversão, foram ignoradas as anotações alternativas e o único atributo extraído das entidades foi a categoria semântica que estão entre 'Abstração', 'Acontecimento', 'Coisa', 'Local', 'Obra', 'Organização', 'Outro', 'Pessoa', 'Tempo' e 'Valor'.

Tabela 1. Estatísticas Primeiro HAREM

Partição	# Sentenças	# Tokens	# Entidades
Partição 1	385	8.836	571
Partição 2	326	7.638	657
Partição 3	419	8.529	406
Partição 4	338	7.559	529
Partição 5	672	13.011	476
Partição 6	207	4.595	403
Partição 7	257	5.493	380
Partição 8	242	4.440	379
Partição 9	392	10.351	669
Partição 10	472	12.149	594
Total	3.710	82.601	5.064

Tabela 2. Estatísticas Mini HAREM

Partição	# Sentenças	# Tokens	# Entidades
Partição 1	303	6.794	388
Partição 2	257	4.898	277
Partição 3	220	5.076	329
Partição 4	294	6.639	361
Partição 5	189	4.415	338
Partição 6	331	5.405	489
Partição 7	179	5.403	330
Partição 8	199	4.861	357
Partição 9	221	5.706	338
Partição 10	273	6.703	462
Total	2.466	55.900	3.669

Tabela 3. Estatísticas Segundo HAREM

Partição	# Sentenças	# Tokens	# Entidades
Partição 1	250	6.150	647
Partição 2	393	8.843	684
Partição 3	396	9.130	736
Partição 4	628	10.679	1.068
Partição 5	470	8.966	953
Partição 6	451	11.081	1.013
Partição 7	286	6.127	616
Partição 8	400	5.488	594
Partição 9	300	8.490	811
Partição 10	267	5.758	695
Total	3.841	80.712	7.817

As Tabela 1, Tabela 2 e Tabela 3 apresentam o número de sentenças, número de

tokens e número de entidades em cada partição gerada para o Primeiro, Mini e Segundo HAREM, respectivamente. As tabelas apresentam também número total de sentenças, tokens e entidades em cada base de dados.

3.2.2. Paramopama

Paramopama [Júnior et al. 2015] é tido como uma base de dados padrão prata (*silver-standard*), sendo construída a partir da melhoria e expansão da versão na língua portuguesa da base WikiNER [Nothman et al. 2013], que possui as seguintes categorias de entidades nomeadas: 'Local', 'Organização', 'Pessoa' e 'Tempo'.

O processo de melhoria das anotações se deu pelo uso de classificadores treinados na base de dados HAREM para rotulação automática de 10.000 sentenças do corpus WikiNER. Posteriormente, a equipe avaliou o processo de rotulação automática, comparando as anotações de todas as sentenças, removendo manualmente as entidades nomeadas rotuladas de forma incorreta.

No processo de expansão, o produto da etapa de melhoria foi utilizado para treinar um classificador e rotular 2.500 novas sentenças, coletadas em sites de notícias com domínios variados, dando preferência para textos contendo entidades do tipo 'Organização' com objetivo de melhoria do desempenho do classificador para tal entidade. O resultado do classificador foi revisado manualmente e adicionado ao corpus final.

Os autores disponibilizaram a base de dados completa em um arquivo texto, sem divisão por documentos. Na etapa de pré-processamento do presente trabalho, as anotações foram convertidas para o formato IOB2, descrito na subseção 3.1.

Tabela 4. Estatísticas Paramopama

Partição	# Sentenças	# Tokens	# Entidades
Partição 1	1.217	31.099	2.336
Partição 2	1.217	30.689	2.386
Partição 3	1.217	30.626	2.325
Partição 4	1.217	30.494	2.377
Partição 5	1.217	30.444	2.380
Partição 6	1.217	30.795	2.335
Partição 7	1.217	30.752	2.356
Partição 8	1.216	30.864	2.384
Partição 9	1.216	30.673	2.333
Partição 10	1.216	30.263	2.411
Total	12.167	306.699	23.623

A Tabela 4 apresenta o número de sentenças, tokens e entidades em cada partição gerada para a base de dados Paramopama. Essa base de dados não possui divisões por documento e, assim, ela foi estratificada a nível de sentença. Isso resultou em uma distribuição mais homogênea de sentenças, tokens e entidades entre as partições quando comparadas com as demais bases de dados.

3.2.3. LeNER-Br

O conjunto de dados LeNER-Br [Luz de Araujo et al. 2018] é considerado o primeiro no domínio de textos legais para a língua portuguesa. Ele é composto por um total de 70 documentos, dos quais 66 são provenientes de tribunais brasileiros e os outros quatro são documentos legislativos, incluindo a Lei Maria da Penha. Assim como os corpus HAREM, os textos do LeNER-Br foram anotados manualmente, tornando-o uma coleção dourada (*golden-collection*). A base de dados possui um total de seis categorias de entidades nomeadas: 'Local', 'Organização', 'Pessoa', 'Tempo', 'Legislação' e 'Jurisprudência', sendo as duas últimas específicas ao domínio de textos legais.

Cada documento do conjunto de dados foi disponibilizado em um arquivo texto, com as anotações seguindo o formato IOB2. Além dos documentos que compõem o conjunto de dados, os autores também forneceram as divisões de treinamento, teste e validação utilizadas para avaliar a performance do conjunto de dados.

Tabela 5. Estatísticas LeNER

Partição	# Sentenças	# Tokens	# Entidades
Partição 1	1.245	37.525	1.632
Partição 2	909	30.730	1.179
Partição 3	304	7.905	305
Partição 4	455	15.089	542
Partição 5	1.913	51.944	1.630
Partição 6	757	23.514	903
Partição 7	1.105	41.086	1.316
Partição 8	1.072	29.166	1.713
Partição 9	2.024	61.681	2.235
Partição 10	616	19.433	793
Total	10.400	318.073	12.248

A Tabela 5 apresenta o número de sentenças, tokens e entidades em cada partição gerada para a base de dados LeNER.

3.2.4. UlyssesNER-Br

Ainda no domínio de textos legais, [Albuquerque et al. 2022] apresentam o corpus UlyssesNER-Br, uma coleção dourada produzida no contexto do Projeto Ulysses, da Câmara dos Deputados Brasileira, que pauta iniciativas institucionais de inteligência artificial visando apoio às atividades legislativas com análises complexas, além de aumento de transparência e melhoria do relacionamento da Câmara com a população.

O UlyssesNER-Br foi dividido em PL-corpus, contendo textos de projetos de lei, e ST-corpus, composto por solicitações de trabalho, documentos internos da Câmara dos

Deputados. Apenas o PL-corpus foi disponibilizado publicamente devido à natureza dos dados utilizados para construção do ST-corpus. Os 150 documentos do PL-corpus foram disponibilizados em arquivos texto com as anotações no formato IOB2. Assim como nas bases de dados HAREM, o corpus UlyssesNER-Br disponibiliza as categorias de entidades nomeadas e a especificação de seus tipos, dividindo as sete categorias semânticas rotuladas em 18 tipos [Albuquerque et al. 2022].

Para o presente trabalho, assim como para os corpus HAREM, foram consideradas apenas as categorias semânticas, que são: 'Data', 'Evento', 'Local', 'Organização', 'Pessoa', 'Fundamento' e 'Produto de Lei', sendo as duas últimas específicas ao domínio da base de dados.

Os autores realizaram a avaliação da base de dados com uma divisão de 75% das sentenças para treino e as 25% restantes para teste. Também foram realizados experimentos de validação cruzada de 5-partes na partição de treino.

Tabela 6. Estatísticas UlyssesNER-Br

Partição	# Sentenças	# Tokens	# Entidades
Partição 1	1.289	15.340	347
Partição 2	705	10.226	317
Partição 3	687	14.121	542
Partição 4	1.022	16.595	384
Partição 5	930	17.661	424
Partição 6	965	13.462	400
Partição 7	545	10.825	284
Partição 8	1.030	10.072	239
Partição 9	1.572	19.123	500
Partição 10	785	11.316	326
Total	9530	138.741	3.763

A Tabela 6 apresenta o número de sentenças, tokens e entidades em cada partição para a base de dados UlyssesNER-Br.

3.2.5. CachacaNER

CachacaNER [Silva et al. 2023] é uma coleção dourada no domínio específico da bebida brasileira cachaça. A base de dados é composta de textos descritivos extraídos de páginas de 24 sites de comércio eletrônico de cachaça. Os dados foram utilizados para analisar as principais categorias presentes, o que resultou em 17 categorias semânticas, das quais 11 são do domínio específico da bebida como Nome da bebida, equipamento de destilação e tempo de armazenamento e seis categorias genéricas como nome de organização e de pessoas.

A base de dados foi dividida em 10 partições com 100 documentos cada, permitindo a realização de experimentos com validação cruzada e facilitando a replicação dos mesmos. Os autores disponibilizaram o corpus em um arquivo csv, onde sentenças foram tokenizadas e cada linha do arquivo representa um token, com os rótulos no formato IOB2 e identificação da sentença, documento e partição à qual o mesmo pertence [Silva et al. 2023].

Tabela 7. Estatísticas CachacaNER

Partição	# Sentenças	# Tokens	# Entidades
Partição 1	1.312	17.290	2.246
Partição 2	1.326	18.793	2.361
Partição 3	1.344	17.954	2.293
Partição 4	1.314	18.280	2.352
Partição 5	1.363	18.347	2.485
Partição 6	1.373	19.085	2.496
Partição 7	1.422	19.631	2.418
Partição 8	1.327	16.749	2.408
Partição 9	1.415	18.052	2.439
Partição 10	1.432	18.838	2.541
Total	13.628	183.019	24.039

A Tabela 7 apresenta o número de sentenças, tokens e entidades em cada partição disponibilizada pelos autores.

4. Avaliação Experimental

A avaliação experimental do trabalho tem com objetivo avaliar o desempenho de um classificador baseado no modelo de linguagem BERTimbau nas sete bases de dados para a tarefa de NER. BERTimbau é um modelo de linguagem baseado em Transformers pré-treinado na língua portuguesa [Souza et al. 2020]. A abordagem proposta pelo presente trabalho busca facilitar a replicação dos experimentos de validação cruzada nessas bases de dados para novos modelos de classificação para NER.

No decorrer desta seção, será descrita a abordagem, as métricas utilizadas e os resultados da avaliação, bem como o ambiente de execução e ferramentas utilizados.

4.1. Configuração Experimental

Os experimentos foram executados em um computador com as seguintes especificações: processador Intel Core i7 com clock de 2.9GHz, 128 GB de memória RAM e uma placa

gráfica NVIDIA GeForce RTX 3090 com 24GB de memória dedicada.

O sistema operacional usado foi o Ubuntu 22.04.2 LTS e os experimentos foram conduzidos utilizando a linguagem de programação Python (versão 3.10.6) juntamente com a biblioteca Transformers (versão 4.25.1), com o framework PyTorch (versão 1.13.1+cu116) como backbone, e a biblioteca Datasets (versão 2.7.1), da plataforma Hugging Face. Além disso, foi empregado também o framework de otimização de hiperparâmetros Optuna (versão 3.1.0) [Akiba et al. 2019].

A abordagem proposta consiste na realização de uma sequência de passos padronizados, sendo eles: (I) Pré-processamento, (II) Ajuste fino e (III) Predição e avaliação.

O processo de ajuste fino (*fine-tuning*) é crucial para a avaliação do desempenho das bases de dados para a tarefa de NER utilizando modelos de linguagem pré-treinados. Em termos gerais, o processo envolve a otimização do modelo na identificação e classificação das entidades nomeadas através do treinamento do mesmo em uma base de dados do domínio da aplicação.

Durante esse processo, o modelo de linguagem pré-treinado é inicializado com os pesos obtidos no seu treinamento em grandes volumes de textos. Esses pesos iniciais fornecem ao modelo um forte entendimento da língua portuguesa e de suas características linguísticas gerais. Esse conhecimento geral é aproveitado na captura de informações contextuais e relações semânticas no texto, o que é essencial para um bom desempenho em tarefas específicas como o NER.

Para garantir robustez e a comparabilidade da avaliação, foi empregada uma estratégia de validação cruzada de 10-partes utilizando as partições disjuntas geradas para cada base de dados. Todos os procedimentos descritos a seguir foram realizados em cada uma das execuções da validação cruzada. O fluxograma na Figura 5 ilustra uma iteração da validação cruzada.

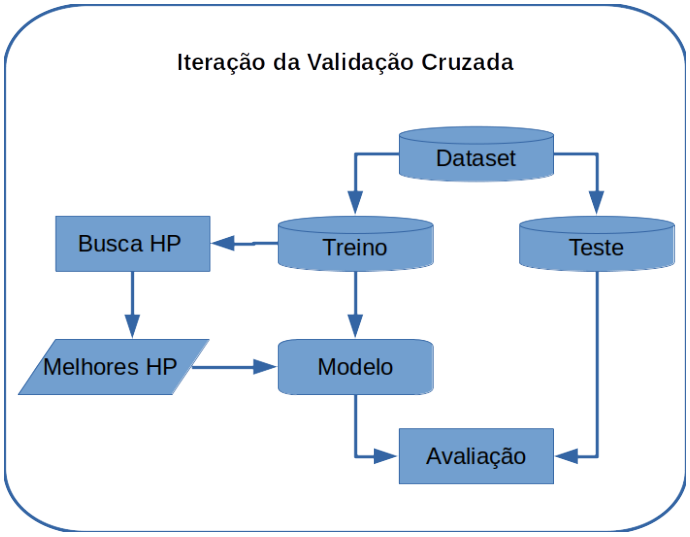


Figura 5. Fluxograma Validação Cruzada

A base de dados foi dividida em conjunto de teste, com uma partição, e conjunto de treino, com as nove partições restantes. Em seguida foi aplicada uma busca de valores para os hiperparâmetros, com intuito de encontrar a melhor configuração para o modelo pré-treinado. A busca foi realizada usando o conjunto de treino, com uma de suas nove partições servindo como conjunto de validação.

Durante a busca foram conduzidos 25 estudos para explorar o espaço de pesquisa. A seleção da melhor combinação foi baseada na maximização da métrica de avaliação F1. Para aprimorar a eficiência da busca, uma estratégia de poda de percentil 25 foi empregada, descartando tentativas (*trials*) pouco promissoras de forma antecipada.

Tabela 8. Espaço de Pesquisa de Hiperparâmetros

Hiperparâmetros	Valores
Learning Rate	1×10^{-6} - 1×10^{-4}
Batch Size	8, 16, 32, 64
Epochs	2, 3, 4, 8, 16, 32, 64, 128
Attention Dropout	0.0 - 0.5
Hidden Dropout	0.0 - 0.5
Weight Decay	0.01 - 0.05

A Tabela 8 apresenta o espaço de pesquisa de cada hiperparâmetro. O espaço abrange os valores indicados por [Devlin et al. 2018] para o processo de ajuste fino do modelo BERT. O valor 64 foi retirado do espaço de pesquisa para o hiperparâmetro *Batch Size* nas bases de dados mini HAREM, LeNER-Br e UlyssesNER-Br devido à restrições de hardware.

Após a busca de hiperparâmetros, o modelo foi treinado no conjunto de treinamento completo usando a configuração ótima obtida na etapa anterior e avaliado no conjunto de teste. Os resultados reportados na Subseção 4.3 representam o desempenho médio do modelo nas 10 iterações da estratégia de validação cruzada.

4.2. Métricas para Avaliação

Foram escolhidas as métricas mais utilizadas na literatura [Silva et al. 2023] [Albuquerque et al. 2023] para a avaliação das bases de dados NER: Revocação (*Recall*), Precisão, Acurácia e F1. O cálculo é realizado a partir da predição a nível de tokens gerada pelo modelo NER, como mostrado a seguir.

$$\text{Revocação} = \frac{TP}{TP + FN} \tag{1}$$

$$\text{Precisão} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$F1 = \frac{2 \times \text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{4}$$

$$Macro\ F1 = \frac{\sum_{i=1}^{n} F1_i}{n} \tag{5}$$

Onde TP são os verdadeiros positivos (*True Positives*); TN, verdadeiros negativos (*True Negatives*); FP, falsos positivos (*False Positives*); FN os falsos negativos (*False Negatives*); e 'n' na equação 5 é o número de categorias.

4.3. Resultados e Discussão

As Tabelas 9, 10, 11, 12 e 13 descrevem o desempenho médio do modelo nas 10 iterações da estratégia de validação cruzada. Os resultados apresentam o desempenho por categoria

de entidade nomeada, bem como as médias macro e micro-F1 (acurácia). As métricas de avaliação predição, revocação e F1 estão representadas por **P**, **R** e **F1** respectivamente.

Tabela 9. Desempenho nas bases de dados HAREM

Categoria	Primeiro			Mini			Segundo		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
ABSTRACCAO	0.612	0.672	0.633	0.537	0.532	0.518	0.582	0.572	0.568
ACONTECIMENTO	0.637	0.521	0.527	0.338	0.348	0.336	0.700	0.604	0.637
COISA	0.809	0.573	0.651	0.474	0.493	0.468	0.711	0.638	0.669
LOCAL	0.799	0.852	0.822	0.792	0.842	0.813	0.806	0.861	0.832
OBRA	0.486	0.582	0.519	0.613	0.665	0.625	0.589	0.685	0.626
ORGANIZACAO	0.709	0.810	0.749	0.713	0.784	0.740	0.747	0.819	0.778
OUTRO	0.244	0.260	0.199	0.000	0.000	0.000	0.512	0.296	0.358
PESSOA	0.814	0.766	0.778	0.833	0.827	0.828	0.864	0.874	0.868
TEMPO	0.905	0.955	0.928	0.850	0.885	0.864	0.859	0.900	0.879
VALOR	0.786	0.818	0.799	0.787	0.824	0.802	0.806	0.820	0.809
Média macro	0.680	0.681	0.661	0.594	0.620	0.599	0.718	0.707	0.702
Acurácia			0.959			0.956			0.952

Analisando a Tabela 9, é possível observar que o desempenho médio do classificador para as categorias ABSTRACCAO, ACONTECIMENTO, COISA, OBRA e OUTRO é inferior às demais categorias em todos os conjuntos HAREM. Uma explicação para tal comportamento é a quantidade pequena de exemplos para tais categorias, sobretudo a categoria OUTRO, que resultou no pior desempenho do modelo em todos conjuntos, chegando a zero para todas as métricas no Mini HAREM.

O classificador alcançou seu melhor desempenho médio na categoria TEMPO em todos os conjuntos. No Primeiro HAREM, essa categoria abrange apenas palavras que representam datas, como “Século XIX”, ou “Maio de 1996”. Em alguns casos o classificador extrapola os limites de entidades da categoria TEMPO, como é o caso do exemplo: “Os acordos concluídos com a Inglaterra em 1642-54-61 estruturaram essa aliança que

marcará profundamente a vida política e econômica de Portugal e do Brasil durante os dois séculos seguintes.” em que o modelo classificou “1642-54-61” como uma só entidade, já na base de dados, os separadores “-” não foram rotulados como TEMPO.

Tabela 10. Desempenho na base de dados Paramopama

Categoria	**P**	**R**	**F1**
LOCAL	0.927	0.939	0.933
ORGANIZACAO	0.854	0.865	0.859
PESSOA	0.934	0.956	0.945
TEMPO	0.803	0.846	0.824
Média macro	0.880	0.902	0.890
Acurácia			0.978

Pela Tabela 10, entre todas as categorias da base de dados Paramopama, o classificador alcançou o melhor desempenho médio nas categorias PESSOA (F1 = 0.945) e LOCAL (F1 = 0.933). Foi identificado um padrão de erro, onde entidades rotuladas como PESSOA foram classificadas como LOCAL pelo modelo, em alguns casos, devido à rotulações incorretas na base de dados, como no exemplo: “Duas pessoas morreram em um acidente ocorrido na noite de ontem no km 2 da rodovia João do Amaral Gurgel, que liga Caçapava a Jambeiro no interior de São Paulo.” onde o modelo classificou “rodovia João do Amaral Gurgel” como LOCAL, enquanto “João do Amaral Gurgel” foi rotulado como PESSOA na base de dados. Na verdade, isso representa uma rotulação incorreta da base de dados.

Ainda sobre as duas categorias, também foram identificadas instâncias onde entidades nomeadas rotuladas como LOCAL foram classificadas como PESSOA, como é o caso do exemplo: “Este pediu formalmente à comunidade internacional que passasse a referir-se ao país como Iran.”, onde a palavra “Iran” foi classificada como PESSOA pelo modelo e rotulada como “LOCAL” na base de dados.

Esse padrão de erro na classificação é mais comum no primeiro caso. É possível que, junto a ambiguidade das entidades entre as categorias, o desbalanceamento entre a

quantidade de exemplos das categorias favoreça mais o primeiro caso.

O classificador teve seu pior desempenho na categoria TEMPO (F1 = 0.824). Essa categoria na base de dados Paramopama abrange diversas expressões temporais, não apenas palavras representando data. Nas análises, não foi encontrado um padrão de erros de classificação entre outras categorias, entretanto, o modelo classificou trechos que não foram rotulados na base de dados, como no exemplo: “No Quênia, quando algum feriado calha no domingo, ele é também comemorado feriado na segunda-feira seguinte.” em que “segunda-feira seguinte” foi classificada como TEMPO pelo modelo, mas não faz parte de nenhuma categoria na base de dados.

Outro tipo de erro identificado foi a classificação de alguns termos precedentes e subsequentes da rotulação na base de dados, como é o caso do seguinte exemplo: “Nos treze meses seguintes Darwin sofreu com sua saúde precária e fez um enorme esforço para escrever um resumo de seu ‘grande livro sobre espécies’.”, onde o modelo classificou como TEMPO o trecho “Nos treze meses seguintes”, enquanto a rotulação da base de dados não possui a palavra “seguintes”.

Em suma, a maior parte dos erros de classificação identificados para a categoria TEMPO estão ligados a inconsistências no padrão de rotulação da base de dados.

Tabela 11. Desempenho na base de dados LeNER-Br

Categoria	**P**	**R**	**F1**
JURISPRUDENCIA	0.805	0.850	0.822
LEGISLACAO	0.895	0.919	0.907
LOCAL	0.837	0.833	0.832
ORGANIZACAO	0.818	0.881	0.844
PESSOA	0.963	0.972	0.967
TEMPO	0.972	0.978	0.975
Média macro	0.882	0.906	0.891
Acurácia			0.979

Na base de dados LeNER-Br, Tabela 11, o classificador alcançou o maior desem-

penho médio nas categorias TEMPO (F1 = 0.975) e PESSOA (F1 = 0.967). Foram poucos os padrões de erros encontrados na classificação da categoria TEMPO, como no seguinte exemplo: “Anota que desde então a recorrida não fez mais a leitura mensal do consumo, permanecendo de dezembro/2016 a abril/2017 sem enviar faturas” onde o modelo classifica todo o trecho “dezembro/2016 a abril/2017” como TEMPO enquanto na base de dados o conectivo “a” não está rotulado.

Tabela 12. Desempenho na base de dados UlyssesNER-Br

Categoria	**P**	**R**	**F1**
DATA	0.959	0.948	0.951
EVENTO	0.513	0.479	0.486
FUNDAMENTO	0.826	0.884	0.852
LOCAL	0.805	0.836	0.813
ORGANIZACAO	0.793	0.838	0.812
PESSOA	0.899	0.899	0.898
PRODUTODELEI	0.628	0.732	0.663
Média macro	0.775	0.802	0.782
Acurácia			0.981

Na base de dados UlyssesNER-Br, Tabela 12 o classificador teve seu maior desempenho na categoria DATA (F1 = 0.951). Dos poucos problemas, foi identificado um padrão em que o modelo tem dificuldade em classificar entidades da categoria DATA que estão entre valores monetários.

O pior desempenho do modelo foi na categoria EVENTO (F1 = 0.486), o que ocorre devido à pequena quantidade de exemplos na base de dados. Analisando os resultados da categoria PRODUTODELEI (F1 = 0.663), foram identificadas situações como: “Recentemente, foi publicada a Lei nº 13.819, de 2016 6, que instituiu a Política Nacional de Prevenção da Automutilação e do Suicídio, e trouxe diversas inovações ao ordenamento jurídico, no contexto da prevenção desse agravo.” em que o modelo classificou o trecho “Política Nacional de Prevenção da Automutilação e Suicídio” uma entidade da

categoria FUNDAMENTO como PRODUTODELEI.

Tabela 13. Desempenho na base de dados CachacaNER

Categoria	P	R	F1
CARACTERISTICA_SENSORIAL_AROMA	0.882	0.874	0.878
CARACTERISTICA_SENSORIAL_CONSISTÊNCIA	0.909	0.946	0.926
CARACTERISTICA_SENSORIAL_COR	0.910	0.931	0.919
CARACTERISTICA_SENSORIAL_SABOR	0.822	0.852	0.836
CLASSIFICACAO_BEBIDA	0.893	0.939	0.914
EQUIPAMENTO_DESTILACAO	0.945	0.957	0.950
GRADUACAO_ALCOOLICA	0.961	0.983	0.972
NOME_BEBIDA	0.914	0.927	0.920
NOME_LOCAL	0.973	0.975	0.974
NOME_ORGANIZACAO	0.888	0.898	0.893
NOME_PESSOA	0.941	0.948	0.944
PRECO	0.875	0.877	0.876
RECIPIENTE_ARMAZENAMENTO	0.958	0.975	0.966
TEMPO	0.963	0.971	0.967
TEMPO_ARMAZENAMENTO	0.980	0.988	0.984
TIPO_MADEIRA	0.976	0.982	0.979
VOLUME	0.991	0.982	0.987
Média macro	0.928	0.941	0.934
Acurácia			0.983

O classificador obteve seu maior desempenho geral na base de dados CachacaNER, como ilustrado na Tabela 13, onde é possível observar que VOLUME foi a categoria com o melhor resultado. Alguns dos poucos erros de classificação dessa categoria estão relacionadas ao limite da entidade nomeada, por exemplo, no trecho “DESCRIÇÃO DA CACHAÇA: Cachaça Ypióca 150 700 ml.” o nome da bebida termina com um valor numérico 150 e é seguido pelo volume (700 ml), e o modelo classificou 150 como parte da entidade VOLUME.

A categoria CARACTERISTICA_SENSORIAL_SABOR (F1 = 0.836) foi a que o classificador teve o seu pior desempenho médio. Isso pode ser explicado pela ambiguidade dos termos utilizados para descrever características sensoriais de sabor e aroma, por exemplo, o termo "ervas" se refere à característica sensorial de sabor no trecho "Encorpada, macia, intensa e persistente, deixa uma boca enxuta, frutada e com forte impressão de ervas.", já no trecho "As cachaças armazenadas em tonéis de bálsamo são conhecidas pelo seu sabor suave e aroma intenso de ervas.", o mesmo termo se refere à característica sensorial de aroma.

5. Conclusão e Trabalhos Futuros

Neste estudo, propusemos uma abordagem para particionar conjuntos de dados utilizados na tarefa de NER, facilitando a reprodutibilidade dos experimentos e possibilitando expansão futura para outros modelos e conjuntos de dados. Na abordagem, o pré-processamento envolveu a conversão dos datasets para o formato IOB2, a quebra dos textos em sentenças e a aplicação de uma estratificação iterativa para criar 10 partições disjuntas. Aplicamos essa metodologia em sete bases de dados na língua portuguesa, com o objetivo de preservar a distribuição das categorias de Entidades Nomeadas presentes nos conjuntos de dados originais. Ademais, realizamos uma avaliação experimental do desempenho de um classificador baseado no modelo de linguagem BERTimbau para cada um dos conjuntos de dados, utilizando a validação cruzada de 10-partes.

No âmbito de trabalhos futuros, sugere-se a avaliação de diferentes modelos de linguagem pré-treinados, além do BERTimbau, a fim de avaliar sua eficácia nas tarefas de Reconhecimento de Entidades Nomeadas em língua portuguesa. Além disso, outras bases de dados NER em português também podem ser particionadas e disponibilizadas usando a mesma metodologia.

Referências

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2623–2631, New York, NY, USA. Association for Computing Machinery.

Albuquerque, H., Souza, E. P., Gomes, C., Pinto, M., Filho, R., Costa, R., Lopes, V., Félix, N., Carvalho, A., and Oliveira, A. (2023). Named entity recognition: a survey for the portuguese language. *Procesamiento de Lenguaje Natural*, 70:171–185.

Albuquerque, H. O., Costa, R., Silvestre, G., Souza, E., da Silva, N. F. F., Vitório, D., Moriyama, G., Martins, L., Soezima, L., Nunes, A., Siqueira, F., Tarrega, J. P., Beinotti, J. V., Dias, M., Silva, M., Gardini, M., Silva, V., de Carvalho, A. C. P. L. F., and Oliveira, A. L. I. (2022). Ulyssesner-br: A corpus of brazilian legislative documents for named entity recognition. In Pinheiro, V., Gamallo, P., Amaro, R., Scarton, C., Batista, F., Silva, D., Magro, C., and Pinto, H., editors, *Computational Processing of the Portuguese Language*, pages 3–14, Cham. Springer International Publishing.

Babych, B. and Hartley, A. (2003). Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT Workshop on MT and Other Language Technology Tools, Improving MT through Other Language Technology Tools: Resources and Tools for Building MT*, EAMT '03, page 1–8, USA. Association for Computational Linguistics.

Consoli, B., Santos, J., Gomes, D., Cordeiro, F., Vieira, R., and Moreira, V. (2020). Embeddings for named entity recognition in geoscience Portuguese literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4625–4630, Marseille, France. European Language Resources Association.

de Lima Santos., D. B., de Carvalho Dutra., F. G., Parreiras., F. S., and Brandão., W. C. (2021). Assessing the effectiveness of multilingual transformer-based text embeddings for named entity recognition in portuguese. In *Proceedings of the 23rd International Conference on Enterprise Information Systems - Volume 1: ICEIS*, pages 473–483. INSTICC, SciTePress.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

do Amaral, D., Collovini, S., Figueira, A., Vieira, R., and Gonzalez, M. (2017). Processo de construção de um corpus anotado com entidades geológicas visando ren. In *Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 63–72. SBC.

Freitas, C., Mota, C., Santos, D., Oliveira, H. G., and Carvalho, P. (2010). Second HAREM: Advancing the state of the art of named entity recognition in Portuguese. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Júnior, C. M., Macedo, H., Bispo, T., Santos, F., Silva, N., and Barbosa, L. (2015). Paramopama: a brazilian-portuguese corpus for named entity recognition. *Encontro Nac. de Int. Artificial e Computacional*.

Luz de Araujo, P. H., de Campos, T. E., de Oliveira, R. R. R., Stauffer, M., Couto, S., and Bermejo, P. (2018). Lener-br: A dataset for named entity recognition in brazilian legal text. In Villavicencio, A., Moreira, V., Abad, A., Caseli, H., Gamallo, P., Ramisch, C., Gonçalo Oliveira, H., and Paetzold, G. H., editors, *Computational Processing of the Portuguese Language*, pages 313–323, Cham. Springer International Publishing.

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., and Gómez-Berbís, J. M. (2013). Named entity recognition: Fallacies, challenges and opportunities. *Computer Standards Interfaces*, 35(5):482–489.

Menezes, D., Milidiu, R., and Savarese, P. (2019). Building a massive corpus for named entity recognition using free open data sources. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 6–11. IEEE.

Mollá, D., Van Zaanen, M., and Smith, D. (2006). Named entity recognition for question answering. In *Proceedings of the Australasian language technology workshop 2006*, pages 51–58.

Nothman, J., Ringland, N., Radford, W., Murphy, T., and Curran, J. R. (2013). Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175. Artificial Intelligence, Wikipedia and Semi-Structured Resources.

Pan, X., Zhang, B., May, J., Nothman, J., Knight, K., and Ji, H. (2017). Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Ramshaw, L. A. and Marcus, M. P. (1999). *Text Chunking Using Transformation-Based Learning*, pages 157–176. Springer Netherlands, Dordrecht.

Santos, D., Seco, N., Cardoso, N., and Vilela, R. (2006). HAREM: An advanced NER evaluation contest for Portuguese. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Sechidis, K., Tsoumakas, G., and Vlahavas, I. (2011). On the stratification of multi-label data. In Gunopulos, D., Hofmann, T., Malerba, D., and Vazirgiannis, M., editors, *Machine Learning and Knowledge Discovery in Databases*, pages 145–158, Berlin, Heidelberg. Springer Berlin Heidelberg.

Silva, P., Franco, A., Santos, T., Brito, M., and Pereira, D. (2023). Cachacaner: a dataset for named entity recognition in texts about the cachaça beverage. *Language Resources and Evaluation*.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Yadav, V. and Bethard, S. (2018). A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.