Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

# Reconhecimento de Entidades Nomeadas com Ensembles de Transformers: Uma Aplicação na Análise de Convênios do Diário Oficial da União

Eutino Junior Vieira Sirqueira

Dissertação apresentada como requisito parcial para
conclusão do Mestrado em Informática

Orientador
Prof. Dr. Flávio de Barros Vidal

Brasília
2026

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

# Reconhecimento de Entidades Nomeadas com Ensembles de Transformers: Uma Aplicação na Análise de Convênios do Diário Oficial da União

Eutino Junior Vieira Sirqueira

Dissertação apresentada como requisito parcial para
conclusão do Mestrado em Informática

Prof. Dr. Flávio de Barros Vidal (Orientador)
CIC/UnB

Prof.a Dr.a Maria Clícia Stelling de Castro
UERJ

Prof.a Dr.a Aleteia Patricia Favacho de Araujo
CIC/UnB

Prof.a Dr.a Cláudia Nalon
Coordenadora do Programa de Pós-graduação em Informática

Brasília, 04 de Fevereiro de 2026

# Dedicatória

*A Deus, fonte de toda a minha força e consolo.*

*À minha esposa, Caroline, meu amor e alicerce diário.*

*À minha irmã, Keli Cristina, pelo incentivo constante.*

*À memória de meus pais, Laurenir e Eutino.*
*Ela, que não me viu começar, mas me ensinou a sonhar.*
*Ele, meu companheiro de batalha, que caminhou comigo até o limiar desta conquista e partiu às vésperas da vitória.*
*A vocês, minha eterna gratidão.*

# Agradecimentos

Agradeço primeiramente a Deus, meu refúgio e fortaleza. Esta conquista é fruto de Sua graça e providência.

À minha família, meu porto seguro. À minha esposa, Caroline, pelo amor incondicional e por ser meu alicerce diário; nossa união foi a razão pela qual não desisti. Esta conquista é tão minha quanto sua.

Ao meu pai, Eutino (in memoriam), cujo exemplo de resiliência e coragem me ensinou sobre a vida mais do que qualquer livro. À memória de minha mãe, Laurenir (in memoriam), minha grande amiga, cuja alegria e generosidade continuam a guiar meus passos.

À minha irmã, Keli Cristina, e ao seu esposo, João Rosa, pelo incentivo e por despertarem em mim a paixão pela Informática. O auxílio deles foi essencial ao longo de toda a minha trajetória.

Ao meu orientador, Prof. Dr. Flávio de Barros Vidal, pela orientação humana, paciência ímpar e por acreditar neste trabalho mesmo diante das adversidades enfrentadas.

À UnB e ao PPGI, pela gestão humanizada e pela sensibilidade institucional, fundamentais para a conclusão desta etapa.

Ao Instituto Federal do Piauí (IFPI). Nesta instituição, fui aluno técnico, servi como técnico de laboratório e hoje tenho a honra de atuar como docente. Agradeço aos colegas de trabalho e amigos pelo incentivo, renovando meu compromisso de retribuir aos alunos tudo o que o IFPI e a educação pública me proporcionaram.

A todos que contribuíram para esta vitória, meu sincero agradecimento.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

*"Passei anos ensinando máquinas a entenderem o contexto das palavras.*
*E, neste processo, Deus me ensinou a entender o contexto da vida:*
*sem paciência e persistência, nenhum modelo converge."*

(Do Autor)

# Resumo

A escassez de *corpora* anotados no domínio de convênios públicos limita o uso de técnicas de Reconhecimento de Entidades Nomeadas (REN) para a fiscalização governamental. Este trabalho aborda essa lacuna desenvolvendo e avaliando modelos baseados na arquitetura *Transformer* e técnicas de *ensemble* para extrair 26 tipos de entidades de documentos do Diário Oficial da União (DOU). Metodologicamente, construiu-se e disponibilizou-se um *corpus* massivo contendo 192.900 documentos e mais de 10,4 milhões de anotações automáticas. Os experimentos compararam a adaptação de sete modelos *Transformer* pré-treinados contra três estratégias de combinação. Os resultados demonstraram que os *ensembles* superaram os modelos individuais, promovendo ganhos significativos de consistência em categorias com menor disponibilidade de dados e maior robustez na correção de erros de omissão. As contribuições incluem a publicação de um conjunto de dados inédito em larga escala e a validação de uma arquitetura eficaz para sistemas de inteligência.

**Palavras-chave:** Reconhecimento de Entidades Nomeadas. Processamento de Linguagem Natural. Modelos *Transformer*. Combinação de Modelos. *Ensemble*. Convênios Públicos.

# Abstract

The scarcity of annotated *corpora* in the public contracts domain limits the application of Named Entity Recognition (NER) techniques for government oversight. This work addresses this gap by developing and evaluating models based on the *Transformer* architecture and *ensemble* techniques to extract 26 entity types from documents of the Official Gazette of the Union (DOU). Methodologically, a massive *corpus* comprising 192,900 documents and over 10.4 million automatic annotations was constructed and made available. The experiments compared the adaptation of seven pre-trained *Transformer* models against three combination strategies. Results demonstrated that the *ensembles* outperformed individual models, promoting significant consistency gains in categories with lower data availability and increased robustness in correcting omission errors. Contributions include the publication of a novel large-scale dataset and the validation of an effective architecture for intelligence systems.

**Keywords:** Named Entity Recognition. Natural Language Processing. Transformer Models. Model Combination. Ensemble Methods. Public Contracts.

# Sumário

# Lista de Figuras

# Lista de Tabelas

# Lista de Abreviaturas e Siglas

**AM** Aprendizado de Máquina.

**AP** Aprendizado Profundo.

**BiLSTM** Bidirectional Long Short-Term Memory.

**CLM** Causal Language Modeling.

**CNN** Convolutional Neural Network.

**CRF** Campos Aleatórios Condicionais.

**DOU** Diário Oficial da União.

**EI** Extração de Informações.

**GPT** Generative Pre-trained Transformer.

**LSTM** Long short-term memory.

**MLM** Masked Language Modeling.

**PLN** Processamento de Linguagem Natural.

**QA** Perguntas e Respostas (do inglês, *Question Answering*).

**REN** Reconhecimento de Entidades Nomeadas.

**RNN** Recurrent Neural Network.

**RNP** Redes Neurais Profundas.

# Capítulo 1

# Introdução

A área de Processamento de Linguagem Natural (PLN) vivenciou avanços significativos nos últimos anos, impulsionados principalmente pelo desenvolvimento de modelos de aprendizado profundo baseados na arquitetura *Transformers* [3, 5]. Dentre as diversas tarefas descritas na literatura, o Reconhecimento de Entidades Nomeadas (REN) — do inglês, *Named Entity Recognition* (NER) — destaca-se por sua ampla aplicabilidade prática [6]. O REN tem como objetivo identificar e classificar entidades relevantes em textos, como nomes de pessoas, organizações, localidades e datas [7]. A sua utilização apresenta grande potencial para a análise e gestão de documentos, especialmente, em contextos legais e governamentais, nos quais a extração precisa de informações é fundamental [8].

No contexto brasileiro, a análise de dados abertos, como os documentos oficiais publicados no Diário Oficial da União (Diário Oficial da União (DOU)) [9], é crucial para a transparência e o acesso à informação [10]. Nesse cenário, o REN se torna uma técnica valiosa quando aplicada a documentos de convênios públicos. O convênio público, conforme o Decreto nº 11.531/2023, é o instrumento que regula a transferência de recursos financeiros da União para a execução de programas e projetos de interesse recíproco [11]. Nesses documentos, a aplicação de REN possibilita a extração automatizada de dados que podem fortalecer iniciativas como a *Deep Vacuity* [12], voltada ao aprimoramento da transparência e da eficiência na gestão pública por meio da consolidação e análise de dados dos Diários Oficiais do Brasil.

Para enfrentar o desafio de analisar o grande volume de dados governamentais foi concebido o projeto *Deep Vacuity*. Conforme descrito em [12], trata-se de uma plataforma que utiliza Aprendizado de Máquina (AM), com base em uma arquitetura de computação de alto desempenho, projetada para coletar, depurar, consolidar e analisar dados publicados nos Diários Oficiais do Brasil. O objetivo da plataforma é oferecer uma interface amigável para auxiliar o trabalho de gestores públicos e órgãos de controle nas funções de fiscalização, auditoria e investigação. Por meio dela, profissionais podem identificar

padrões, tendências e, principalmente, indícios de inconformidades ou irregularidades de forma mais eficiente, apoiando a fiscalização e promovendo uma gestão pública mais eficaz e transparente.

Apesar do sucesso em analisar documentos em larga escala, a plataforma *Deep Vacuity* enfrenta o desafio de extrair informações de forma estruturada dos textos, que são por natureza heterogêneos. É neste ponto que o REN se torna uma tecnologia essencial. A aplicação de modelos de REN permite transformar o conteúdo textual dos Diários Oficiais em dados organizados, identificando e categorizando entidades como órgãos públicos, empresas, valores e datas — elementos fundamentais para o monitoramento de convênios.

Este trabalho, portanto, contribui diretamente para o projeto *Deep Vacuity* ao desenvolver e disponibilizar um componente especializado de REN voltado a documentos de convênios. O objetivo central é viabilizar a extração automatizada de informações críticas, enriquecendo a base de dados da plataforma e fornecendo aos tomadores de decisão informações mais granulares e estruturadas para suas análises.

Apesar do potencial do REN no contexto apresentado, sua aplicação prática na análise de documentos públicos enfrenta desafios significativos que precisam ser superados. A complexidade da tarefa pode ser resumida em quatro obstáculos principais:

1. **Volume e Velocidade dos Dados:** O Diário Oficial da União publica um volume massivo de documentos diariamente. Qualquer solução de análise precisa ser automatizada e escalável para lidar com essa produção contínua de dados não estruturados;

2. **Variabilidade e Ambiguidade Textual:** Os extratos de convênios, embora sigam padrões, apresentam alta variabilidade na redação. Além disso, a ambiguidade é um desafio constante; por exemplo, um termo como "Fundo Nacional de Desenvolvimento"pode, dependendo do contexto, se referir ao `NOME ÓRGÃO CONCEDENTE` ou ao `NOME ÓRGÃO SUPERIOR`;

3. **Ausência de um *Corpus* Anotado:** O principal obstáculo técnico para o treinamento de modelos de AM supervisionados é a inexistência de um *corpus* de referência, público e em larga escala, para o domínio específico de convênios públicos no Brasil. Sua criação, portanto, tornou-se um pré-requisito para a pesquisa;

4. **Necessidade de Alto Desempenho:** Para que a extração automática seja útil em cenários práticos, como os do projeto *Deep Vacuity*, os modelos precisam atingir um alto grau de acurácia e completude, minimizando a necessidade de revisão manual e garantindo a confiabilidade das informações.

Para superar tais desafios, uma abordagem promissora é o uso de combinação de modelos (*ensemble*), que busca agregar as "decisões"de diversos modelos para produzir um

resultado final mais robusto do que qualquer um de seus componentes isoladamente [13]. Nesse contexto, este trabalho estabeleceu os objetivos descritos na Seção 1.1.

## 1.1 Objetivos

O objetivo geral deste trabalho é investigar e avaliar o impacto de técnicas de *ensemble* [14], aplicadas a um comitê de modelos baseados na arquitetura *Transformer* [3], como estratégia para aprimorar o desempenho e a robustez do Reconhecimento de Entidades Nomeadas em extratos de convênios públicos do DOU.

Para alcançar o objetivo geral, os seguintes objetivos específicos foram traçados e cumpridos:

1. Construir um *corpus* em larga escala para o domínio de convênios públicos por meio de uma estratégia de anotação automática;
2. Avaliar o desempenho de sete diferentes modelos baseados na arquitetura *Transformer* para a tarefa de REN, após o processo de ajuste fino (*fine-tuning*);
3. Comparar o desempenho de três estratégias de *ensemble* (Votação do Melhor Modelo, Votação por Maioria e Votação Ponderada) aplicadas aos modelos treinados;
4. Analisar os resultados quantitativos e qualitativos, com foco no comportamento dos modelos e *ensembles* perante o desbalanceamento de classes e categorias de entidades.

## 1.2 Contribuições e Organização do Documento

Para atender aos objetivos traçados, este trabalho apresenta contribuições metodológicas e práticas significativas. A principal delas, desenvolvida em colaboração com Peres [4], foi a criação de um *corpus* inédito para o domínio de convênios públicos, composto por 192.900 publicações do DOU, totalizando 10.434.116 entidades anotadas distribuídas em 26 classes.

A partir dessa base de dados, foi aplicada a técnica de Transferência de Aprendizado (*Transfer Learning*) para ajustar sete modelos pré-treinados da arquitetura *Transformer* [3], que foram posteriormente combinados por meio de três estratégias de *ensemble*. A validade e a relevância desta abordagem experimental foram reconhecidas pela comunidade científica, resultando na publicação e apresentação de artigo completo nos anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC) [15], conferindo respaldo técnico aos resultados aqui apresentados.

A ausência de trabalhos focados neste contexto específico e a necessidade de colaborar com projetos estratégicos, como o *Deep Vacuity*, motivaram a criação do *corpus* e a avaliação detalhada das estratégias de combinação. Dessa forma, este estudo contribui não apenas para o avanço das técnicas de REN no Brasil, mas também para iniciativas voltadas à eficiência na gestão de documentos públicos.

O restante desta dissertação está estruturado em cinco capítulos. O Capítulo 2 estabelece as bases teóricas da pesquisa, abordando o contexto dos convênios na administração pública, os modelos *Transformers* e os métodos de *ensemble*. O Capítulo 3 revisa o estado da arte e posiciona o trabalho em relação à literatura existente. O Capítulo 4 descreve a arquitetura experimental, detalhando as etapas de construção do *corpus* anotado e a implementação das estratégias de combinação de modelos *Transformers*. O Capítulo 5 apresenta a análise dos experimentos, discutindo o desempenho dos modelos e suas combinações sob perspectivas quantitativas e qualitativas. Por fim, o Capítulo 6 sintetiza as contribuições da pesquisa, suas limitações e aponta caminhos para trabalhos futuros.

# Capítulo 2

# Fundamentação Teórica

Este capítulo apresenta a fundamentação teórica que sustenta a pesquisa desenvolvida. O objetivo é fornecer ao leitor os conceitos essenciais para a compreensão do domínio de aplicação, da metodologia adotada e da correta interpretação dos resultados discutidos, posteriormente, no Capítulo 5.

A discussão inicia na Seção 2.1, a qual contextualiza os convênios na administração pública federal, caracterizando o domínio textual sobre o qual o *corpus* desta pesquisa foi construído. Em seguida, a Seção 2.2 introduz o campo do Processamento de Linguagem Natural (PLN), servindo de base para a definição da tarefa central deste trabalho, o Reconhecimento de Entidades Nomeadas (REN), detalhada na Seção 2.3.

Para fundamentar a análise experimental, a Seção 2.4 apresenta os métodos e as métricas utilizados para aferir o desempenho dos sistemas. Na sequência, a Seção 2.5 traça a evolução das abordagens para REN, partindo dos métodos baseados em regras e em Aprendizado de Máquina tradicional até as arquiteturas de Aprendizado Profundo, com ênfase especial nos modelos baseados em *Transformers*, que representam o estado da arte atual.

Por fim, o capítulo aprofunda-se na estratégia principal investigada neste trabalho, a combinação de modelos. A Seção 2.6 detalha a teoria dos *Ensembles*, discutindo o equilíbrio entre viés e variância, bem como as estratégias de votação aplicadas para maximizar o desempenho dos modelos no domínio jurídico-administrativo.

## 2.1 Convênios na Administração Pública

No âmbito da Administração Pública Federal, o convênio constitui o instrumento jurídico que formaliza a cooperação entre entes federados ou entidades privadas para a realização de objetivos comuns. Historicamente regido pelo art. 116 da Lei nº 8.666/1993 [16] e, atualmente, pela Lei nº 14.133/2021 [17] e pelo Decreto nº 11.531/2023 [11], este

instrumento viabiliza a descentralização da execução de políticas públicas mediante a transferência de recursos da União para a execução de obras, a aquisição de equipamentos ou a realização de eventos [18].

Para compreender a estrutura textual do *corpus* deste trabalho, é fundamental distinguir a natureza jurídica do objeto de estudo. Conforme explicado no trabalho [19], o convênio diferencia-se dos contratos administrativos tradicionais pela convergência de vontades. Enquanto, no contrato, os interesses são opostos e contraditórios (a Administração deseja a obra e o contratado almeja a remuneração), no convênio, os partícipes possuem interesses comuns e coincidentes, caracterizados pela busca de colaboração mútua.

No contexto da Ciência de Dados e do PLN, os documentos de convênios publicados no DOU apresentam desafios específicos que dificultam a automação. Embora a legislação estabeleça os elementos obrigatórios da publicidade, a ausência de um padrão rígido transforma o que deveria ser um dado estruturado em um texto livre e ruidoso.

Um ponto crucial que justifica a extração direta do DOU é a tempestividade da informação. Conforme declarado pelo próprio Portal da Transparência do Governo Federal [20], a atualização dos dados estruturados ocorre com periodicidade mensal. Em contrapartida, o DOU é publicado diariamente. Além disso, o Portal frequentemente reflete o estado *atual* do convênio (incluindo termos aditivos de valor e prazo), enquanto a extração do DOU permite recuperar o *histórico* dos atos originais. A Figura 2.1 ilustra um caso real (Convênio nº 886044/2019) que evidencia essas discrepâncias e desafios:

- **Divergência Temporal e de Valores:** O texto original do DOU (B) registra o valor de R$ 1.977.197,00 e a vigência até 2021. O dado estruturado no Portal (A) já apresenta o valor atualizado de R$ 2.217.038,44 e vigência até 2025. Sem a extração do texto do DOU, perde-se o rastro da pactuação original;
- **Densidade de Entidades:** O parágrafo do extrato condensa, em poucas linhas, múltiplas entidades nomeadas (Organizações, Datas, Valores Monetários) que precisam ser desambiguadas;
- **Variação de Redação:** O objeto do convênio no DOU é apresentado de forma resumida ou ligeiramente diferente do cadastro oficial, o que exige modelos robustos de linguagem.

Portanto, a aplicação de técnicas de REN neste domínio não visa apenas a extração de texto, mas também à estruturação de dados governamentais essenciais para a auditoria em tempo real e a preservação da memória administrativa.

**A) Dado Estruturado (Fonte: Portal da Transparência - Estado Atual):**

```
{
   "Numero_Convenio":  "886044",
   "Concedente":  "UNIVERSIDADE FEDERAL RURAL DO SEMI-ARIDO - RN",
   "Convenente":  "FUNDACAO GUIMARAES DUQUE",
   "Valor_Celebrado":  "R$ 2.217.038,44", // Valor atualizado c/ aditivos
   "Fim_Vigencia":  "19/03/2025"      // Data estendida
}
```

---

**B) Texto Não Estruturado (Fonte: DOU de 11/12/2019 - Ato Original):**

*"**EXTRATO DE CONVÊNIO**. Espécie: Convênio 03/2019, Nº Plataforma Mais Brasil: 886044/2019. Concedente: Universidade Federal Rural do Semi-Árido, CNPJ: 24.529.265/0001-40; Convenente: Fundação Guimarães Duque, CNPJ: 08.350.241/0001-72. Objeto: Apoio administrativo [...] ao projeto "Desenvolvimento de conteúdo técnico [...]". Valor Total: **R$ 1.977.197,00**. [...] Vigência: 12/12/2019 a **12/12/2021**. Data de assinatura: 10/12/2019. [...]"*

Figura 2.1: Comparativo entre o dado estruturado (atualizado) e o texto original publicado no DOU. Nota-se que o extrato textual preserva os valores e prazos originais (R$ 1,9 mi; 2021), enquanto a base estruturada reflete atualizações posteriores (R$ 2,2 mi; 2025), evidenciando a importância da extração do DOU para a auditoria do ato originário.

## 2.2 Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) é uma subárea da Inteligência Artificial e da Ciência da Computação cujo objetivo é habilitar os computadores a compreender, interpretar, gerar e manipular a linguagem humana [21]. Historicamente, essa tarefa apresenta desafios significativos devido à natureza inerentemente ambígua, contextual e não estruturada da comunicação humana, o que torna complexa sua representação em formatos processáveis por máquinas [21].

Contudo, um ponto de inflexão na área foi o advento e a consolidação do Aprendizado Profundo (do inglês, *Deep Learning*), que, conforme aponta [22], tem gerado resultados extremamente promissores em tarefas de compreensão da linguagem. De acordo com [23], a evolução para Redes Neurais Profundas (RNP), baseada em representações vetoriais densas, produziu um salto de desempenho em uma vasta gama de aplicações de PLN. Entre as tarefas que mais se beneficiaram, destacam-se:

- Modelagem de linguagem (do inglês, *language modeling*) [24];
- Etiquetagem de classes gramaticais (do inglês, *POS tagging*) [25];
- Reconhecimento de Entidades Nomeadas (do inglês, *Named Entity Recognition*) [26];
- Análise sintática (do inglês, *parsing*) [27];

- Etiquetagem de papéis semânticos (do inglês, *Semantic Role Labeling*) [28];
- Classificação de texto e análise de sentimentos (do inglês, *text and sentiment classification*) [29];
- Sumarização de texto (do inglês, *summarization*) [30];
- Tradução automática (do inglês, *machine translation*) [31];
- Resposta a perguntas (do inglês, *question answering*) [32];
- Recuperação de informação (do inglês, *information retrieval*) [33].

Esses avanços demonstram a crescente capacidade computacional para lidar com a complexidade da linguagem humana, impulsionada pela sofisticação de técnicas como as RNP [22, 23]. Tais técnicas não apenas aprimoraram o desempenho em tarefas já consolidadas, mas também ampliaram o nível de complexidade das aplicações possíveis no campo do PLN [23]. Dentre essa vasta gama de tarefas, o Reconhecimento de Entidades Nomeadas (REN) assume particular relevância neste trabalho e é explorado em detalhe a seguir.

## 2.3 Reconhecimento de Entidades Nomeadas (REN)

O Reconhecimento de Entidades Nomeadas (REN), do inglês *Named Entity Recognition* (NER), é uma tarefa fundamental do Processamento de Linguagem Natural (PLN) que consiste em localizar e classificar menções de entidades em textos não estruturados, associando-as a categorias predefinidas, como nomes de pessoas, organizações e locais [1, 34, 35]. O termo "Entidade Nomeada" foi formalizado em 1996, durante a Sexta Conferência de Compreensão de Mensagens (do inglês, *Sixth Message Understanding Conference* (MUC-6)), um evento focado em Extração de Informações (EI) de textos jornalísticos [36, 37]. Naquela conferência, as entidades foram definidas para incluir não apenas nomes próprios, mas também expressões numéricas como datas, horários, valores monetários e porcentagens.

Desde a concepção inicial, o escopo da tarefa foi ampliado. Atualmente, conforme aponta [1], as entidades podem ser categorizadas em domínios gerais (como nomes de organizações, pessoas e locais) ou em domínios específicos. No domínio biomédico, por exemplo, o REN é utilizado para identificar nomes de genes, proteínas, drogas e doenças, demonstrando a versatilidade e a importância da tarefa [38]. Essencialmente, o processo de REN envolve duas etapas: a detecção dos limites da entidade no texto, que pode ser composta por uma ou mais palavras (*tokens*), e a posterior classificação dessa entidade em uma das categorias predefinidas.

Do ponto de vista formal, dada uma sequência de *tokens* $s = (w_1, w_2, \ldots, w_N)$, uma solução de REN deve gerar uma lista de tuplas $(I_{start}, I_{end}, t)$, em que cada tupla representa uma entidade nomeada em $s$ [1]. Nessa definição, $I_{start}$ e $I_{end}$ são os índices que delimitam a posição da entidade na sequência, e $t$ é o tipo da entidade, pertencente a um conjunto de categorias predefinido.

Na prática, para treinar modelos de Aprendizado de Máquina, essa tarefa é comumente modelada como um problema de etiquetagem de sequência (*sequence tagging*)[39]. A abordagem mais difundida para isso é a utilização de um esquema de anotação como o **BIO**, que significa ***B**eginning* (Início), ***I**nside* (Dentro) e ***O**utside* (Fora) [35]. Nesse esquema, cada *token* na sequência de entrada recebe uma etiqueta que indica sua posição em relação a uma entidade:

- **B-TIPO**: marca o *token* que **inicia** uma entidade do tipo TIPO;
- **I-TIPO**: marca os *tokens* que estão **dentro** (continuação) de uma entidade do tipo TIPO;
- **O**: marca os *tokens* que estão **fora** de qualquer entidade.

Por exemplo, para a sentença "O Ministério da Educação fica em Brasília.", com as entidades `NOME ÓRGÃO CONCEDENTE` e `LOCAL`, a etiquetagem no formato BIO seria:

```
O           O
Ministério  B-NOME_ORGAO_CONCEDENTE
da          I-NOME_ORGAO_CONCEDENTE
Educação    I-NOME_ORGAO_CONCEDENTE
fica        O
em          O
Brasília    B-LOCAL
.           O
```

Essa representação permite que modelos de Aprendizado Profundo, como os *Transformers*, aprendam a prever a etiqueta correta para cada *token* da sequência, resolvendo, assim, a tarefa de REN.

A Figura 2.2 apresenta outro exemplo genérico de um REN. O processo parte de uma sequência de palavras $S = (w_1, w_2, \ldots, w_n)$, como a frase "Eutino Júnior nasceu em Corrente, Piauí". O modelo de REN identifica entidades na sequência com base em tipos predefinidos, como Pessoa e Local, atribuindo etiquetas às palavras ou conjuntos de palavras que correspondem a essas categorias. No exemplo, as entidades "Eutino Júnior" são classificadas como Pessoa ($< w_1, w_2, Pessoa >$), enquanto "Corrente" e "Piauí" são identificadas como Locais ($< w_5, w_5, Local >$ e $< w_6, w_6, Local >$, respectivamente). Esse

processo ilustra a capacidade do REN de segmentar e categorizar informações em um texto.

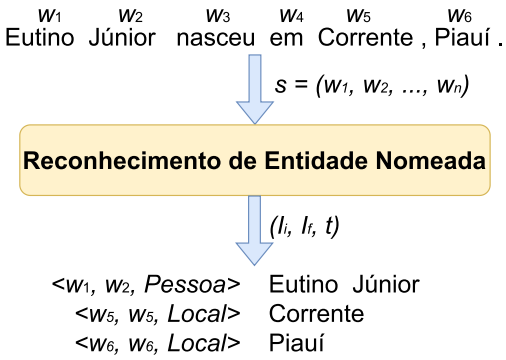



Figura 2.2: Exemplo de REN em uma frase. Fonte: Adaptado de [1].

É importante notar que, além do esquema BIO, que opera no nível dos *tokens*, existem outras formas de representar as entidades. Uma abordagem alternativa, adotada por diversas ferramentas modernas como a biblioteca *spaCy* [40], é a anotação baseada em *spans*. Neste formato, em vez de etiquetar cada *token*, especificam-se diretamente os índices de início e fim da entidade, em nível de caracteres, no texto original e foi o formato adotado para o desenvolvimento deste trabalho.

### 2.3.1 Domínios de Aplicação do REN

A versatilidade do REN é evidenciada por sua aplicação em uma vasta gama de domínios, que vão desde a análise de mídias digitais até o aprimoramento de sistemas complexos. Alguns exemplos, são:

- **Monitoramento de Mídias Sociais.** Neste domínio, o REN é fundamental para identificar automaticamente menções a marcas, empresas ou pessoas, apoiando análises de mercado, percepção de sentimento e gestão de crises. A tecnologia permite, ainda, a detecção de tendências e a avaliação do impacto de campanhas digitais, como destacam os trabalhos [41] e [42];

  **Auxílio à Desambiguação de Entidades.** O REN é um pré-requisito essencial para a tarefa de desambiguação de entidades, um desafio que ocorre quando um mesmo termo possui múltiplos significados (e.g., *Apple* como empresa ou fruta) [35]. Após a identificação da entidade pelo REN, técnicas como as propostas por [43], [44] e [45] são aplicadas para associá-la ao seu sentido correto no contexto;

- **Sistemas de Perguntas e Respostas (*Question Answering*).** Em sistemas de Perguntas e Respostas (do inglês, *Question Answering*) (QA), o REN desempenha

um papel crucial ao identificar as entidades-chave nas perguntas dos usuários (e.g., "Qual a capital da França?" ). Essa identificação permite que o sistema restrinja a busca por respostas a documentos ou passagens que contenham a entidade "França" , tornando a recuperação de informações factuais significativamente mais eficiente, como demonstrado em [46] e [47];

- **Tradução Automática.** Na tradução automática, o REN contribui para a qualidade do resultado final ao garantir que nomes próprios e outras entidades sejam corretamente identificados e preservados ou transliterados, em vez de serem traduzidos erroneamente como palavras comuns [35, 48, 49].

A diversidade destas aplicações demonstra que o REN se consolidou como uma tecnologia transversal e de alto impacto. Para o contexto deste trabalho, que foca na análise de convênios do Diário Oficial da União, essas experiências validam o potencial do REN para processar e estruturar grandes volumes de dados governamentais. Isso viabiliza não apenas a extração de informações, mas também a criação de bases de conhecimento que podem ser utilizadas para aprimorar a transparência e a eficiência na administração pública.

## 2.4 Métodos de Avaliação de Sistemas de Reconhecimento de Entidades Nomeadas

A avaliação de desempenho de um sistema de REN é um processo fundamental que consiste em comparar as entidades previstas pelo modelo com um gabarito de referência — geralmente um *corpus* anotado por especialistas humanos, que representa a "verdade fundamental" (*ground truth*) [1, 35]. Historicamente, essa comparação é realizada por meio de duas abordagens principais: a avaliação exata (*exact match*) e a avaliação relaxada (*relaxed match*) [38]. As conferências de avaliação competitiva, como o CoNLL-2002 [50] e o CoNLL-2003 [51], foram cruciais para consolidar a avaliação exata como o padrão na área. A distinção entre as duas abordagens está nos critérios utilizados para considerar uma predição como correta:

- A avaliação exata, mais rígida e comum, exige que uma entidade prevista seja considerada correta somente se ambos os seus limites (início e fim) e o seu tipo coincidirem perfeitamente com a entidade no gabarito. Qualquer desvio, seja no limite ou no tipo, classifica a predição como incorreta;
- A avaliação relaxada, por sua vez, adota critérios mais flexíveis e pode atribuir crédito parcial a previsões que não sejam perfeitas. Por exemplo, uma predição

pode ser considerada parcialmente correta se acertar os limites, mas errar o tipo, ou vice-versa. Essa abordagem foi utilizada em padrões de avaliação como MUC [36] e ACE [52].

É importante destacar que a escolha das ferramentas utilizadas neste trabalho está alinhada à abordagem mais rigorosa. O treinamento e a avaliação dos modelos foram realizados utilizando a *framework* spaCy, uma biblioteca de PLN amplamente adotada na indústria e na academia [40]. A sua ferramenta de avaliação nativa implementa o critério de avaliação exata [40], seguindo o padrão estabelecido pela conferência CoNLL-2003 [51]. Dessa forma, todos os resultados apresentados neste trabalho consideram uma entidade correta apenas quando há correspondência perfeita de limites e de tipo com o gabarito, garantindo uma análise de desempenho rigorosa e comparável a outros estudos da área. As métricas quantitativas utilizadas para medir o desempenho sob esse critério são detalhadas a seguir.

### 2.4.1 Métricas de Avaliação

Para quantificar o desempenho de um sistema de REN sob o critério de avaliação exata, de acordo com [1, 35, 37, 38, 51], utilizam-se três métricas padrão, derivadas da área de Recuperação de Informação: Precisão (*Precision*), Revocação (*Recall*) e Medida-F1 (*F1-Score*). O cálculo dessas métricas depende da comparação entre as predições de um modelo e as anotações de referência da verdade fundamental (*ground truth*). Conforme a literatura [35, 53], cada *token* ou sequência de *tokens* classificada pode resultar em um de quatro possíveis resultados:

- **Verdadeiro Positivo (VP):** Ocorre quando o modelo identifica corretamente uma entidade, acertando tanto seus limites (início e fim) quanto seu tipo. Este é o resultado desejado;
- **Falso Positivo (FP):** Ocorre quando o modelo aponta uma entidade que não consta no gabarito. Trata-se de um erro de comissão, pois o modelo executou uma ação de classificação que não deveria ter sido feita;
- **Falso Negativo (FN):** Ocorre quando o modelo falha em identificar uma entidade presente no gabarito. Trata-se de um erro de omissão, pois o modelo deixou de executar uma ação de classificação que era necessária;
- **Verdadeiro Negativo (VN):** Refere-se a um *token* que não faz parte de uma entidade e foi corretamente classificado como "não-entidade" (geralmente com a etiqueta O no esquema BIO). Embora seja um acerto, esta medida não é diretamente

utilizada no cálculo das métricas padrão de REN, pois a quantidade de verdadeiros negativos (a grande maioria dos *tokens* em um texto) é tão elevada que ofuscaria a avaliação do desempenho nas classes de entidades que são, de fato, o objeto de interesse.

Com base nesses componentes e conforme [35], as métricas são definidas da seguinte forma. A **Precisão** mede a proporção de entidades que o modelo previu corretamente em relação a todas as que ele previu. Em outras palavras, de tudo o que o sistema rotulou, quanto estava correto. O seu cálculo é dado pela Equação 2.1.

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{2.1}$$

A **Revocação** (*Recall*) mede a proporção de entidades relevantes do gabarito que o modelo identificou. Ou seja, de tudo o que deveria ser encontrado, quanto o sistema de fato encontrou. O seu cálculo é dado pela Equação 2.2.

$$\text{Revocação} = \frac{VP}{VP + FN} \tag{2.2}$$

A Precisão e a Revocação são métricas que frequentemente apresentam uma relação inversa, um fenômeno conhecido na literatura como o dilema ou ***trade-off*** **Precisão-Revocação** [53]. Em geral, um modelo configurado para ser mais cauteloso pode atingir alta precisão ao custo de uma baixa revocação, pois só classifica as entidades sobre as quais tem alta certeza. Inversamente, um modelo mais liberal pode alcançar alta revocação, identificando a maioria das entidades, mas ao custo de uma baixa precisão, pois comete mais erros.

Para obter uma medida única que equilibre essas duas métricas, a abordagem padrão é utilizar a **F1-Score** (ou Medida-F1). Por ser a média harmônica entre a Precisão e a Revocação, a *F1-Score* penaliza valores extremos e, portanto, constitui uma métrica robusta para avaliar o desempenho geral e equilibrado de um modelo [1]. O cálculo dessa métrica é apresentado na Equação 2.3.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{2.3}$$

## 2.5 Evolução das Abordagens para Reconhecimento de Entidades Nomeadas

Esta seção apresenta uma perspectiva sobre a evolução das abordagens para REN. A Figura 2.3 ilustra o panorama dessas abordagens, detalhadas sequencialmente, desde os

métodos clássicos baseados em regras até as arquiteturas modernas de Aprendizado Profundo e *Transformers*.

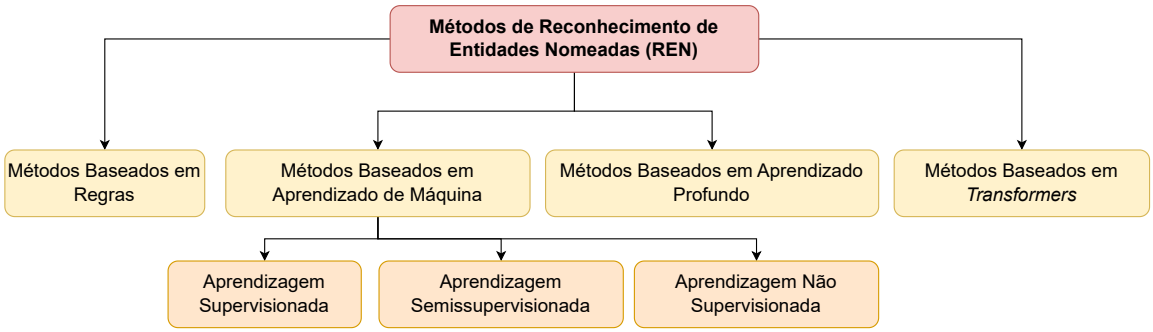



Figura 2.3: Panorama das abordagens para REN, desde sistemas baseados em regras até modelos baseados em *Transformers*.

As abordagens para o Reconhecimento de Entidades Nomeadas apresentadas na figura 2.3 serão detalhadas individualmente nas seções subsequentes.

### 2.5.1 Métodos Baseados em Regras

Os métodos baseados em regras representam a abordagem clássica e uma das primeiras estratégias para o REN. Esses sistemas operam com base em conhecimento linguístico e em um domínio codificado manualmente, sendo tipicamente estruturados em três componentes principais [35]: (1) um conjunto de regras que utiliza marcadores para identificar entidades; (2) dicionários (ou *gazetteers*) que compilam termos de um domínio; e (3) um mecanismo de extração que aplica as regras e os dicionários ao texto de entrada. As regras, por sua vez, podem explorar padrões morfológicos e sequências de palavras (regras baseadas em padrões) ou depender de pistas semânticas e do entorno linguístico da menção (regras baseadas em contexto) [54].

O principal ponto forte dessa abordagem é a capacidade de alcançar alta **Precisão** em domínios bem delimitados, uma vez que as regras são definidas por especialistas para minimizar falsos positivos [35]. Contudo, essa precisão frequentemente vem ao custo de uma baixa **Revocação** [1]. A cobertura restrita dos dicionários e a rigidez das regras fazem com que muitas entidades válidas não sejam reconhecidas, o que resulta em um alto número de falsos negativos. Além disso, esses sistemas sofrem de baixa portabilidade e de alto custo de manutenção [55, 56]. A adaptação para novos domínios ou idiomas exige a criação manual de novas regras e vocabulários, um processo trabalhoso, de difícil escalabilidade e que depende de conhecimento especializado [34].

Apesar de suas limitações, os métodos baseados em regras ainda são relevantes em cenários específicos. São uma alternativa viável em situações com escassez de dados

anotados, onde o treinamento de modelos de Aprendizado de Máquina é impraticável, ou quando se deseja incorporar conhecimento de um especialista de forma explícita e controlada no sistema [35]. Exemplos na literatura demonstram sua eficácia, como no trabalho de [57] sobre a língua *Urdu*, que superou modelos estatísticos da época devido à falta de um *corpus* de qualidade. Seu uso continua em estudos atuais que buscam alta precisão em nichos específicos [58, 59].

Em síntese, a abordagem baseada em regras oferece controle e alta precisão em domínios restritos, mas sua rigidez e alto custo de manutenção limitam sua aplicação em cenários mais amplos e dinâmicos. Essas limitações, especialmente a dificuldade de generalização e adaptação, motivaram a transição para abordagens orientadas a dados, baseadas em Aprendizado de Máquina, que são discutidas a seguir.

### 2.5.2 Métodos Baseados em Aprendizado de Máquina

As limitações dos sistemas baseados em regras, especialmente a dificuldade de adaptação e o alto custo de manutenção, motivaram a transição para abordagens orientadas a dados, como o Aprendizado de Máquina (AM). Conforme definido por [60], AM é um campo da Inteligência Artificial focado no desenvolvimento de algoritmos capazes de aprender padrões a partir de grandes volumes de dados, sem serem programados explicitamente para cada tarefa.

No contexto do REN, isso significa treinar um modelo estatístico para que ele aprenda a tomar decisões de classificação de entidades com base em exemplos. A forma como esses exemplos são utilizados define o paradigma de aprendizado, que pode ser supervisionado, semi-supervisionado ou não supervisionado. As técnicas mais recentes, baseadas em Aprendizado Profundo, representam uma evolução natural desses métodos e são exploradas na Seção 2.5.3.

#### Aprendizagem Supervisionada

A Aprendizagem Supervisionada é o paradigma dominante no REN clássico. Nessa abordagem, o modelo aprende a partir de um *corpus* de treinamento previamente rotulado por especialistas, que contém exemplos de entidades em seus respectivos contextos [55]. A premissa é que, ao analisar esses exemplos, o modelo consiga generalizar o conhecimento para classificar entidades em textos novos e nunca vistos. A qualidade e a quantidade dos dados anotados são, portanto, fatores críticos que determinam diretamente o desempenho e a capacidade de generalização do modelo final [35, 37].

Um processo central nos métodos supervisionados tradicionais é a engenharia de características (*feature engineering*). Esse processo consiste em projetar e extrair manualmente

atributos informativos do texto que ajudem o modelo a distinguir entre entidades e não-entidades [1]. Essas características criam uma representação vetorial para cada palavra, que pode incluir uma vasta gama de informações, tais como [55, 61]:

- **Características léxicas e ortográficas:** A própria palavra, seu lema, se está em maiúsculas, se contém hífens ou dígitos;
- **Características morfológicas:** Prefixos e sufixos da palavra;
- **Recursos externos:** Informações de dicionários, como listas de locais (*gazetteers*) ou de nomes de empresas;
- **Características de contexto:** As palavras e suas características em uma janela de vizinhança.

Com base nessas características, diversos algoritmos foram aplicados ao REN, incluindo Modelos Ocultos de Markov (HMM) [62, 63], Modelos de Máxima Entropia (ME) [64, 65] e Máquinas de Vetores de Suporte (SVM) [66, 67]. Dentre eles, os **Campos Aleatórios Condicionais** (*Conditional Random Fields* - CRF) se destacam como o modelo de maior sucesso para tarefas de etiquetagem de sequência como o REN, por sua capacidade de considerar o contexto global da sentença ao fazer predições [68, 69, 70].

Em suma, embora os métodos supervisionados clássicos possam alcançar bons resultados, sua forte dependência da custosa anotação manual de dados e da complexa engenharia de características motivou a busca por abordagens alternativas.

### Aprendizagem Semi-supervisionado

A Aprendizagem Semi-supervisionado surge como uma solução para mitigar o principal gargalo da abordagem supervisionada: a escassez de dados rotulados. A estratégia central consiste em utilizar um pequeno conjunto de dados rotulados, em conjunto com um grande volume de dados não rotulados, para aprimorar o desempenho do modelo [35, 71].

Uma das técnicas conhecidas nesse paradigma é o *bootstrapping* [55]. O processo geralmente se inicia por treinar um classificador inicial com os poucos dados rotulados disponíveis. Esse classificador é então utilizado para rotular o *corpus* não rotulado. As predições com maior grau de confiança são selecionadas e adicionadas ao conjunto de treinamento original. O modelo é então retreinado com este conjunto de dados ampliado e o processo se repete iterativamente, "ensinando" o modelo a partir de seus próprios acertos [72].

Ao aproveitar a informação contida nos dados não rotulados, esses métodos podem reduzir a dependência da anotação manual. No entanto, enfrentam desafios próprios, principalmente o risco de propagação de erros: se o modelo rotula incorretamente um

dado e o adiciona ao conjunto de treinamento, esse erro pode ser amplificado nas iterações seguintes, degradando o desempenho do sistema [35, 73].

#### Aprendizagem Não Supervisionada

A Aprendizagem Não Supervisionada representa a abordagem para o REN em cenários em que não há dados rotulados disponíveis. O objetivo é descobrir padrões e estruturas latentes diretamente a partir do texto bruto, geralmente por meio de técnicas de agrupamento (*clustering*) [60, 74]. A premissa fundamental é que menções a entidades do mesmo tipo tendem a ocorrer em contextos similares [1].

Esses métodos operam agrupando sintagmas ou unidades linguísticas com base em características compartilhadas, como similaridade de contexto, distribuições de palavras ou funções sintáticas [35]. Por exemplo, um algoritmo pode observar que os termos "Google", "IBM" e "Microsoft" frequentemente aparecem após verbos como "anunciou" ou antes de termos como "Inc.", agrupando-os em um mesmo *cluster* que, para um humano, corresponderia à classe "ORGANIZAÇÃO" [75, 76].

A principal vantagem dessa abordagem é a eliminação total da necessidade de anotações manuais. Contudo, suas limitações são significativas. A principal delas é a dificuldade de atribuir rótulos semânticos significativos aos *clusters* gerados (o modelo pode agrupar empresas, mas não saberá chamá-las de "ORGANIZAÇÃO"). Além disso, a avaliação do desempenho é desafiadora na ausência de uma verdade fundamental, e os modelos podem ter dificuldade em desambiguar termos com múltiplos significados [35].

### 2.5.3 Métodos Baseados em Aprendizado Profundo

A transição para o Aprendizado Profundo (AP) representou um marco na evolução do REN, levando a saltos de desempenho que estabeleceram um novo estado da arte [7]. A principal vantagem dessa abordagem em relação ao Aprendizado de Máquina clássico reside em sua capacidade de aprender representações e características relevantes diretamente dos dados brutos, eliminando a necessidade da complexa e trabalhosa etapa de engenharia de características manuais [22]. Utilizando Redes Neurais Profundas (RNP) com múltiplas camadas, os modelos de AP são capazes de descobrir automaticamente padrões hierárquicos e estruturas latentes em textos, o que os torna especialmente eficazes para lidar com a complexidade da linguagem natural [1].

A arquitetura padrão para a maioria dos modelos de REN baseados em Aprendizado Profundo (na era pré-Transformers) consolidou-se em um *pipeline* de três estágios principais, conforme descrito por [35]: (1) Representação de Dados, onde o texto de entrada é convertido em vetores numéricos; (2) Codificação de Contexto, responsável por capturar

as dependências sequenciais entre os *tokens*; e (3) Decodificação de Entidade, que produz a sequência final de etiquetas. Cada um desses estágios é detalhado a seguir.

### Representação de Dados

A primeira etapa em um modelo de AP para REN consiste em converter o texto em uma representação vetorial que a rede neural possa processar. Para isso, utilizam-se técnicas de *embedding* que transformam unidades linguísticas (palavras ou caracteres) em vetores densos em um espaço de alta dimensão [35]. Duas importantes abordagens são:

- **Embeddings de Palavras (*Word Embeddings*):** Convertem cada palavra do vocabulário em um vetor. Diferente de representações esparsas como *one-hot encoding*, técnicas como *Word2Vec* [77], *GloVe* [78] e *fastText* [79] aprendem representações densas a partir de grandes volumes de texto, capturando relações semânticas e sintáticas (e.g., o vetor de "rei" está para "rainha" assim como "homem" está para "mulher");
- **Embeddings de Caracteres (*Character Embeddings*):** Representam as palavras a partir de seus caracteres constituintes. Essa abordagem é eficaz para lidar com palavras raras ou fora do vocabulário de treinamento (*out-of-vocabulary*) e para capturar informações morfológicas, como prefixos e sufixos [35]. As sequências de caracteres de uma palavra são geralmente processadas por uma Convolutional Neural Network (CNN) ou uma Recurrent Neural Network (RNN) (como uma Bidirectional Long Short-Term Memory (BiLSTM)) para gerar um vetor final da palavra [80, 81].

Na prática, muitos modelos de ponta combinam ambos os tipos de *embeddings* para aproveitar tanto a informação semântica no nível da palavra, quanto a morfológica no nível do caractere [71, 82].

### Codificação de Contexto

Após a conversão do texto em vetores, a camada de codificação de contexto tem o objetivo de processar a sequência de *embeddings* para capturar as dependências entre as palavras e enriquecer a representação de cada *token* com informações de seu entorno [35]. Duas importantes arquiteturas utilizadas para essa finalidade na era pré-Transformers foram [**?**]:

- **Redes Neurais Convolucionais (CNN):** Embora originalmente concebidas para visão computacional, as CNNs foram adaptadas para tarefas de PLN para extrair características locais, como n-gramas de caracteres ou palavras [35, 83]. Elas aplicam "filtros" que deslizam sobre a sequência de *embeddings* para capturar padrões

morfológicos e locais, sendo eficazes para a extração de características de baixo nível [83];

- **Redes Neurais Recorrentes (RNN):** As RNNs são arquiteturas naturalmente adequadas para processar dados sequenciais. Variantes mais robustas, como a *Long short-term memory (LSTM)* [84] e a *Gated Recurrent Unit* (GRU) [85], superam as limitações das RNNs simples, sendo capazes de capturar dependências de longo prazo. Para REN, a arquitetura mais dominante foi a **BiLSTM**, que processa a sequência de entrada em ambas as direções (do início para o fim e do fim para o início), permitindo que a representação de cada palavra contenha informações contextuais de toda a sentença [35, 71, 39].

### Decodificação de Entidade

A etapa final, a decodificação, recebe as representações contextuais da camada anterior e produz a sequência de etiquetas de REN (p. ex., 'B-PESSOA', 'I-PESSOA', 'O'). O principal desafio aqui é garantir que a sequência de etiquetas seja válida e coerente (por exemplo, uma etiqueta 'I-PESSOA' não deve vir após uma etiqueta 'B-LOCAL'). As duas abordagens mais comuns para esta camada são [35]:

- **Camada Softmax:** Trata a etiquetagem de cada *token* como um problema de classificação multiclasse independente. Uma função Softmax é aplicada à saída do codificador para cada *token*, gerando uma distribuição de probabilidade sobre todas as etiquetas possíveis [61]. Embora simples e rápida, essa abordagem não considera as dependências entre as etiquetas vizinhas, podendo gerar sequências inválidas [35];
- **Campos Aleatórios Condicionais (CRF):** Em vez de classificar cada *token* isoladamente, a camada CRF considera a sequência de saídas como um todo [35]. Ela adiciona uma matriz de pesos que aprende as probabilidades de transição entre as etiquetas (por exemplo, a probabilidade de uma etiqueta 'I-PESSOA' seguir uma 'B-PESSOA' é alta). Ao final, utiliza-se um algoritmo de decodificação, como o **algoritmo de Viterbi** [86], para encontrar a sequência de etiquetas com a maior pontuação global, garantindo a coerência do resultado. A combinação de uma camada de codificação BiLSTM com uma camada de decodificação CRF (**BiLSTM-CRF**) tornou-se a arquitetura de facto para REN antes da ascensão dos *Transformers* [71, 82].

A evolução contínua dessas arquiteturas culminou no desenvolvimento dos *Transformers* [3], que introduziram uma nova forma de capturar o contexto e revolucionaram o estado da arte em PLN, como detalhado na próxima seção.

### 2.5.4 Métodos Baseados em *Transformers*

Conforme observado em [2], a história da era moderna do PLN começa em 2017, com a publicação do artigo "*Attention Is All You Need*" [3], que introduziu a arquitetura *Transformer*. No entanto, foi em 2018 que o paradigma do pré-treinamento em larga escala foi estabelecido por dois trabalhos seminais que definiram as duas principais linhagens de arquiteturas: o Generative Pre-trained Transformer (GPT) [87], pioneiro no pré-treinamento de decodificadores para geração de texto, e o BERT [88], que demonstrou o poder dos codificadores bidirecionais para tarefas de compreensão de linguagem. Esses dois modelos deram início a uma rápida sucessão de inovações [2].

O ano de 2019 viu uma explosão de novos modelos que refinaram e expandiram essas ideias. Foi lançado o GPT-2 [89], uma versão ampliada e mais capaz do GPT original. No mesmo ano, surgiram o DistilBERT [90], uma versão menor e mais rápida do BERT que mantinha grande parte de seu desempenho, e modelos *sequence-to-sequence* como o BART [91] e o T5 [92], que unificaram diversas tarefas de PLN em um único formato.

A corrida por escalabilidade e por um melhor alinhamento com a intenção humana marcaram os anos seguintes [2]. Em 2020, o GPT-3 [93] demonstrou capacidades impressionantes de aprendizado com poucos exemplos (*few-shot learning*) e até mesmo sem exemplos (*zero-shot*). Em 2022, o InstructGPT [94] (base do modelo ChatGPT original) introduziu o treinamento com *feedback* humano para tornar os modelos mais úteis e seguros ao seguir instruções.

A partir de 2023, a competição e a inovação se intensificaram com o lançamento de modelos de grande porte de código aberto, como o LLaMA [95] e sua sucessão, e o Mistral 7B [96], que demonstrou um desempenho notável para seu tamanho relativamente compacto. Mais recentemente, em 2024, a pesquisa explorou tanto os limites da escala quanto a eficiência em modelos menores [2]. O Gemma [97] foi lançado como uma família de modelos abertos de alto desempenho, e o SmolLM [98] demonstrou que arquiteturas extremamente compactas podem alcançar resultados competitivos, viabilizando aplicações em dispositivos com recursos limitados.

Essa rápida evolução resultou em um ecossistema diversificado de modelos que, apesar de suas particularidades, podem ser organizados em três famílias arquitetônicas principais [2]: (1) modelos autorregressivos (estilo GPT), baseados no decodificador; (2) modelos autocodificadores (estilo BERT), baseados no codificador; e (3) modelos sequência a sequência (estilo T5/BART), que utilizam a arquitetura completa [2]. Essas famílias são exploradas em mais detalhes adiante.

### *Transformers* como Modelos de Linguagem

Os modelos apresentados na seção anterior, como BERT e GPT, são desenvolvidos como Modelos de Linguagem (*Language Models*). Isso significa que, antes de serem aplicados a qualquer tarefa específica, eles passam por uma fase de pré-treinamento sobre volumes massivos de texto não rotulado (e.g., a Wikipédia, coleções de livros, etc.) [2]. O aprendizado ocorre de forma autossupervisionada, ou seja, o modelo gera seus próprios rótulos a partir dos dados de entrada, dispensando a necessidade de anotação humana nesta etapa inicial [2].

Essa fase de pré-treinamento permite que o modelo desenvolva uma profunda compreensão estatística da linguagem, aprendendo gramática, sintaxe, semântica e até mesmo certo grau de conhecimento de mundo. Existem duas estratégias de pré-treinamento principais que deram origem às diferentes famílias de *Transformers*, conforme detalhado a seguir [2].

**Modelagem de Linguagem Causal (*Causal Language Modeling (CLM)*).** Esta estratégia treina o modelo para prever a próxima palavra em uma sequência, tendo visto apenas as palavras anteriores. É uma abordagem autorregressiva, que processa o texto em uma única direção (geralmente da esquerda para a direita), como ilustrado na Figura 2.4. Este objetivo de treinamento é a base dos modelos da família GPT [87], os quais os tornam especialistas em tarefas de geração de texto.

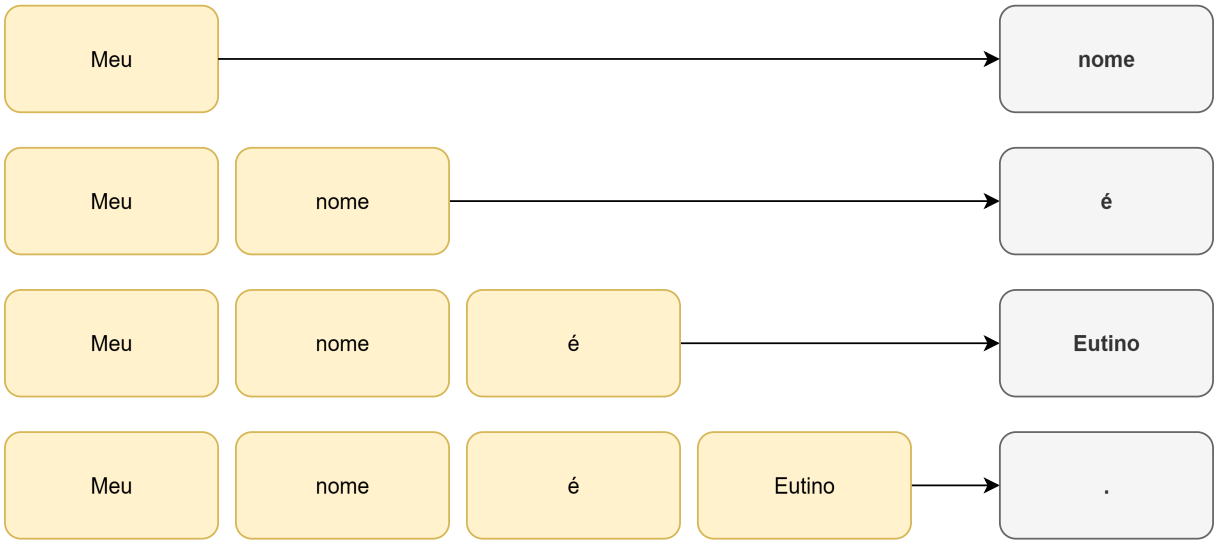



Figura 2.4: Ilustração da Modelagem de Linguagem Causal (CLM), no qual o modelo prevê o próximo *token* da sequência. Adaptado de [2].

**Modelagem de Linguagem Mascarada (*Masked Language Modeling (MLM)*).** Diferente da CLM, a estratégia de MLM consiste em corromper a sentença de entrada,

mascarando (ocultando) aleatoriamente alguns de seus *tokens*, e então treinar o modelo para prever quais eram os *tokens* originais. A Figura 2.5 ilustra este processo. Como o modelo precisa considerar o contexto de ambos os lados (esquerdo e direito) para fazer uma previsão coerente, ele aprende representações textuais bidirecionais profundas. Este objetivo de treinamento é a inovação central do BERT [88] e o que torna os modelos da família *encoder-only* tão poderosos em tarefas de compreensão de sentenças, como o REN.

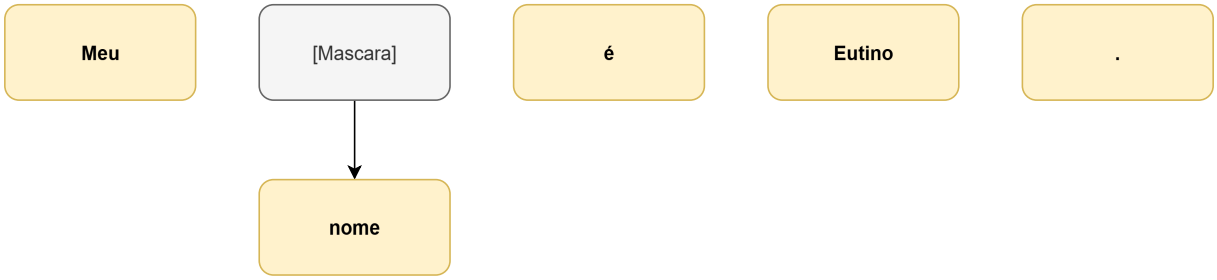



Figura 2.5: Ilustração da Modelagem de Linguagem Mascarada (MLM), no qual o modelo preenche os *tokens* que foram ocultados. Adaptado de [2].

O conhecimento linguístico adquirido durante o pré-treinamento, por si só, não é diretamente aplicável a tarefas práticas como o REN. Para isso, é necessária uma segunda etapa de especialização, realizada por meio da Transferência de Aprendizagem, detalhada a seguir.

### O Paradigma de Transferência de Aprendizagem

A eficácia dos modelos *Transformers* modernos não reside apenas em sua arquitetura, mas no paradigma de Transferência de Aprendizagem (*Transfer Learning*) sob o qual são desenvolvidos [99]. Também conhecida como ajuste fino (*fine-tuning*) ou adaptação de domínio, esta abordagem consiste em duas fases distintas: pré-treinamento e ajuste fino [1].

A primeira fase, o **pré-treinamento**, consiste em treinar um modelo a partir do zero com base em um *corpus* massivo e genérico, como ilustra a Figura 2.6. Esse processo exige grande poder computacional e dias ou semanas de treinamento, resultando em um modelo de linguagem com vasto conhecimento estatístico da língua, mas ainda não especializado para uma tarefa prática [2].

A segunda fase, o **ajuste fino** (*fine-tuning*), ocorre em seguida. Neste processo, o modelo já pré-treinado é alimentado com um novo conjunto de dados, menor e específico da tarefa-alvo (e.g., um *corpus* de REN), como mostrado na Figura 2.7. O treinamento continua por um período muito mais curto, ajustando os pesos do modelo para o novo

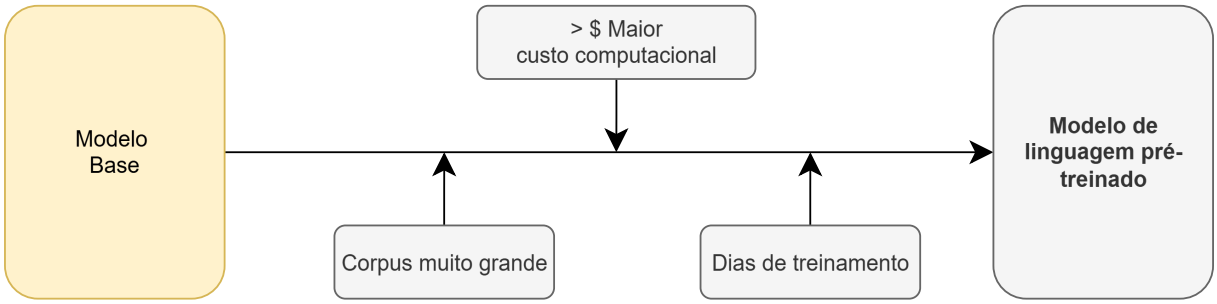



Figura 2.6: Ideia geral do pré-treinamento de um modelo *Transformer*. Adaptado de [2].

domínio [60].Essa abordagem de duas fases apresenta vantagens significativas, tornando o uso de grandes modelos viável e eficaz [2]:

- O modelo já possui um conhecimento linguístico robusto, facilitando a adaptação ao novo contexto;
- A quantidade de dados específicos da tarefa necessária para o ajuste fino é muito menor;
- O custo computacional, o tempo de treinamento e o impacto ambiental são drasticamente reduzidos em comparação ao pré-treinamento;
- Permite maior agilidade para testar e especializar modelos para diferentes aplicações.

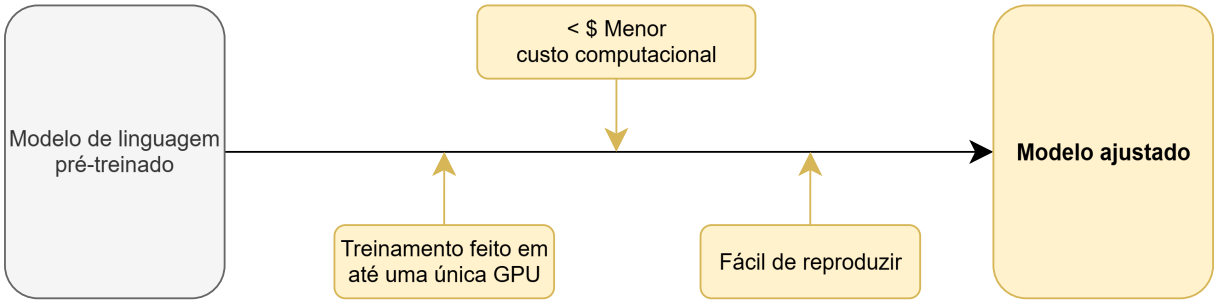



Figura 2.7: Ideia geral do ajuste fino de um modelo *Transformer*. Adaptado de [2].

Em suma, o paradigma de transferência de aprendizagem é o que torna os grandes modelos de linguagem práticos e acessíveis. Ele permite que o conhecimento adquirido de vastos corpora seja "transferido" e especializado para tarefas de nicho , como o reconhecimento de entidades em convênios públicos, com muito menos dados e recursos. Tendo entendido este processo, a próxima seção detalha os componentes internos da arquitetura *Transformer* que possibilitam esse aprendizado.

### A Arquitetura *Transformer* em Detalhe

Conforme introduzido por [3], a principal inovação da arquitetura *Transformer* é a sua capacidade de processar sequências inteiras de uma só vez, baseando-se exclusivamente em mecanismos de atenção para capturar as relações entre os *tokens*, dispensando totalmente a recorrência das RNNs. A arquitetura canônica é composta por um **codificador** (*encoder*) e um **decodificador** (*decoder*), conforme ilustrado na Figura 2.8. Cada bloco do codificador e do decodificador é composto por subcamadas que executam funções específicas, sendo a mais importante o mecanismo de atenção.

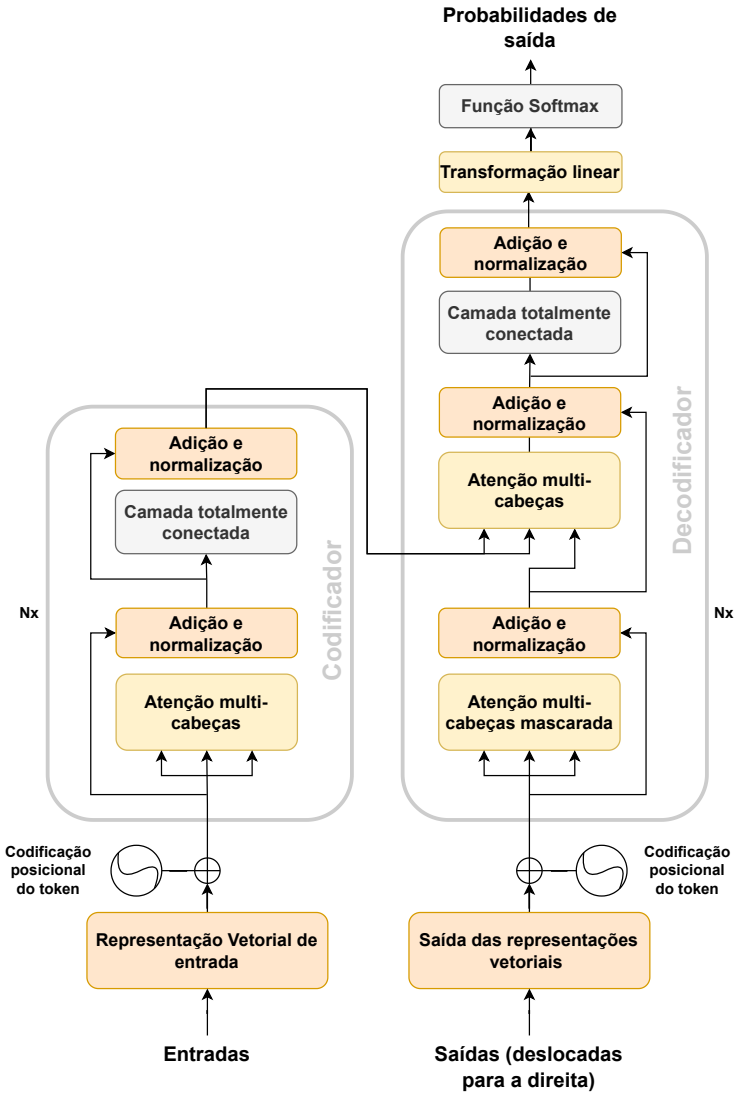



Figura 2.8: A arquitetura *Transformer*, destacando os blocos de Codificador e Decodificador. Fonte: Adaptado de [3].

**O Mecanismo de Atenção Multi-cabeça (*Multi-Head Attention*).** Esta é a peça central do *Transformer*. Conforme visto em [3], a sua forma mais básica, a autoatenção, permite que o modelo, ao processar uma palavra, pondere a importância de todas as demais palavras na mesma sequência. Isso é feito por meio de três vetores, um para cada *token* de entrada: a **Consulta** (*Query*), a **Chave** (*Key*) e o **Valor** (*Value*). A pontuação de atenção é calculada pela compatibilidade entre a *Query* de um *token* e as *Keys* de todos os demais. Essa pontuação determina o peso que o *Value* de cada *token* terá na representação final desse *token* inicial. O *Transformer* aprimora isso com a atenção multicabeça, que realiza esse processo em paralelo múltiplas vezes, permitindo que o modelo capture diferentes tipos de relações contextuais simultaneamente [3].

A Figura 2.9 ilustra o cálculo exato da atenção, conhecido como Atenção com Produto Escalar Escalonado (*Scaled Dot-Product Attention*). O fluxo consiste em calcular o produto escalar entre as *Queries* (Q) e as *Keys* (K), escalonar o resultado para estabilizar o treinamento (dividindo pela raiz quadrada da dimensão das chaves, $\sqrt{d_k}$), aplicar opcionalmente uma máscara e converter os escores em probabilidades com a função *Softmax*. Por fim, essas probabilidades são aplicadas aos *Values* (V) para gerar a representação de saída.

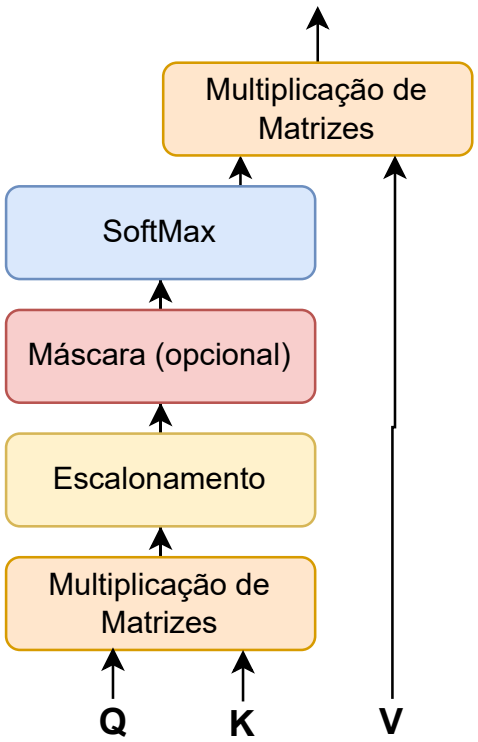



Figura 2.9: O mecanismo de Atenção com Produto Escalar Escalonado. Fonte: Adaptado de [3].

**Componentes Adicionais.** Além da atenção, cada camada do *Transformer* contém outros dois componentes essenciais [3]: uma Rede Neural *Feed-Forward* totalmente conectada, que aplica uma transformação não-linear à saída da camada de atenção, e **conexões residuais** seguidas de normalização de camada (*Layer Normalization*) ao redor de cada subcamada. Esses dois últimos elementos são cruciais para permitir o treinamento de modelos muito profundos, garantindo que o gradiente flua adequadamente e estabilizando o processo de aprendizado [3].

**Impacto e Vantagens.** Nos experimentos originais, o *Transformer* demonstrou um desempenho superior ao dos modelos baseados em RNN e CNN em tarefas de tradução automática, atingindo um novo estado da arte no desafio WMT 2014 com uma pontuação de 28,4 BLEU [3]. A principal vantagem que permitiu essa revolução foi sua alta capacidade de paralelização. Como cada palavra na sequência é processada simultaneamente, em vez de uma após a outra como nas RNNs, o treinamento pôde ser drasticamente acelerado e escalado para volumes de dados e tamanhos de modelo antes impraticáveis [3]. Essa eficiência de treinamento foi o que abriu caminho para a era dos grandes modelos de linguagem pré-treinados, como BERT e GPT [87, 88].

## 2.6 Combinação de Modelos (*Ensembles*)

Após a análise das arquiteturas de Aprendizado Profundo e *Transformers*, esta seção aborda a etapa final da estratégia proposta neste trabalho: a combinação de modelos, ou *Ensemble Learning*. Um *ensemble* é uma abordagem de AM na qual múltiplos modelos de aprendizado, conhecidos como aprendizes base (*base learners*), são estrategicamente combinados para produzir uma predição final que é mais precisa e robusta do que qualquer um dos modelos individuais conseguiria isoladamente [13, 100]. A premissa fundamental é que, ao combinar as previsões de diversos modelos, as falhas específicas de um modelo tendem a ser compensadas pelos acertos de outros, resultando em uma melhor generalização [101]. A base teórica para a combinação dos modelos também passa pela manipulação do viés e da variância, conceitos que também são apresentados a seguir.

### 2.6.1 Fundamentos Teóricos: O Equilíbrio entre Viés e Variância

A eficácia dos métodos de *ensemble* pode ser formalmente compreendida por meio do dilema viés-variância (*bias-variance trade-off*), um conceito central em AM [102]. O erro

de generalização de um modelo pode ser decomposto em três componentes: viés, variância e erro irredutível [102], os quais são:

- **Viés (*Bias*):** Refere-se ao erro decorrente de suposições incorretas no algoritmo de aprendizado. Um alto viés pode fazer com que o modelo não capture as relações relevantes entre as características e a saída (subajuste ou *underfitting*);
- **Variância (*Variance*):** Refere-se à sensibilidade do modelo a pequenas variações no conjunto de treinamento. Uma alta variância pode fazer com que o modelo capture ruído aleatório em vez do sinal pretendido (sobreajuste ou *overfitting*).

Os métodos de *ensemble* atuam diretamente na redução de um ou ambos os componentes do erro. Técnicas como *Bagging* são, em sua maioria, redutoras de variância, enquanto as como *Boosting* focam na redução do viés [101].

### 2.6.2 Principais Tipos de Métodos de *Ensemble*

As estratégias de *ensemble* podem ser classificadas em várias famílias principais, sendo as mais influentes o *Bagging*, o *Boosting* e o *Stacking* [103]. Cada uma dessas categorias fundamenta-se em diferentes lógicas de treinamento e agregação, as quais são detalhadas nos tópicos a seguir.

***Bagging (Bootstrap Aggregating).*** Introduzido por [104], o *Bagging* consiste em treinar múltiplos modelos do mesmo tipo, de forma independente e em paralelo, sobre diferentes subconjuntos de dados criados por amostragem com reposição (*bootstrap*). As predições finais são então combinadas, geralmente por meio de votação majoritária (para classificação) ou de média ponderada (para regressão). Ao treinar os modelos em diferentes visões dos dados, o *Bagging* reduz a variância do modelo final. O algoritmo *Random Forest* [105] é o exemplo mais canônico de *Bagging*, em que os aprendizes base são árvores de decisão.

***Boosting.*** Diferente do *Bagging*, o *Boosting* treina os modelos de forma sequencial e adaptativa. Cada novo modelo na sequência é treinado para dar mais atenção aos exemplos classificados incorretamente pelos modelos anteriores [106]. O objetivo é transformar um conjunto de "aprendizes fracos" (modelos com desempenho ligeiramente melhor que o aleatório) em um único "aprendiz forte". Algoritmos como *AdaBoost* [106] e, mais notavelmente, *Gradient Boosting* [107] são exemplos que focam principalmente na redução do viés.

***Stacking.*** Proposto por [108], o *Stacking* (ou empilhamento) realiza a combinação de modelos de uma forma diferente. Em vez de usar uma função simples (como média ou votação) para combinar as predições, o *Stacking* treina um novo modelo, chamado de **meta-modelo**, para aprender a combinar as saídas dos modelos base. As predições dos modelos base servem como características de entrada (*features*) para o meta-modelo, que então produz a predição final.

### 2.6.3 Estratégias de Combinação por Votação

A votação consiste em uma das formas mais simples e eficazes de combinar as predições de múltiplos modelos treinados de forma independente [14]. A intuição por trás dessa técnica é que, se um número significativo de modelos robustos concorda com uma predição, há uma maior probabilidade de que ela esteja correta [14]. Entretanto, os princípios de votação fundamentam-se em diferentes lógicas; as duas estratégias que serviram de base para este trabalho são detalhadas nos tópicos a seguir.

**Votação Majoritária (*Hard Voting*).** Nesta abordagem, a predição final é a classe que recebe o maior número de votos dos modelos individuais [14]. O princípio conceitual para uma tarefa de classificação é selecionar a moda das predições, conforme formalizado na Equação 2.4:

$$F(x) = \text{mode}\{f_1(x), f_2(x), \ldots, f_N(x)\} \tag{2.4}$$

onde $f_i(x)$ é a predição de classe do i-ésimo modelo para a entrada $x$.

**Votação Ponderada (*Soft Voting*).** Nesta abordagem, em vez de uma única classe, cada modelo fornece um vetor de probabilidades que indica sua "confiança" em relação a cada classe possível. A predição final é a classe que maximiza a soma (ou média ponderada) das probabilidades previstas por todos os modelos. Essa abordagem é geralmente preferida, pois aproveita a informação de confiança de cada modelo [14]. O princípio é expresso pela Equação 2.5:

$$F(x) = \text{argmax}_{c \in C} \sum_{i=1}^{N} w_i p_{i,c}(x) \tag{2.5}$$

onde $p_{i,c}(x)$ representa a probabilidade prevista pelo modelo $i$ para a classe $c$, e $w_i$ é um peso atribuído ao i-ésimo modelo.

Estes dois princípios de votação servem de base conceitual para uma variedade de estratégias de combinação. No entanto, a aplicação direta a tarefas como o REN exige adaptações, uma vez que a predição ocorre no nível do *token* e não no da sentença inteira.

As variações específicas desenvolvidas e implementadas neste trabalho, baseadas nesses conceitos de votação para combinar as previsões dos sete modelos de *Transformers*, são detalhadas no Capítulo 4 da Metodologia.

### 2.6.4 Condições para o Sucesso e a Importância da Diversidade

A literatura de Aprendizado de Máquina estabelece a Acurácia e a Diversidade dos modelos como condições essenciais para que um método de *ensemble* seja eficaz e supere o desempenho de seus componentes individuais [101, 103]. O critério de Acurácia e o de Diversidade são apresentados a seguir.

1. **Acurácia:** Os modelos base devem ter um desempenho, em média, superior ao de um palpite aleatório. Se os modelos individuais não forem informativos, a combinação deles também não será;
2. **Diversidade:** Os modelos base devem ser diversos, ou seja, devem cometer erros em exemplos distintos. Se todos os modelos acertam e erram exatamente nos mesmos lugares, a combinação não trará benefício, pois a votação majoritária, por exemplo, sempre resultará na mesma resposta do modelo individual.

A diversidade pode ser alcançada treinando modelos com diferentes arquiteturas, diferentes hiperparâmetros, ou sobre diferentes subconjuntos de dados, como é o caso deste trabalho, que, conforme visto na Tabela 4.5, utilizou sete variantes da arquitetura *Transformer* pré-treinadas.

### 2.6.5 Aplicação de Ensembles em *Transformers* para PLN

No contexto de modelos de PLN modernos, a aplicação de *ensembles* tem se mostrado uma estratégia robusta para obter ganhos de desempenho, que podem ser significativos em aplicações críticas [109]. A abordagem capitaliza o fato de que, mesmo baseados na mesma arquitetura, diferentes modelos pré-treinados (e.g., BERT, RoBERTa, ELECTRA) aprendem representações ligeiramente distintas do mesmo texto. Essa diversidade sutil resulta em erros não correlacionados, que podem ser mitigados por meio da combinação de suas previsões [109].

A eficácia dessa estratégia é demonstrada em diversos contextos linguísticos e em domínios de alta especialização. Para o Reconhecimento de Entidades Nomeadas, trabalhos mostram que a combinação de múltiplos modelos *Transformer* resulta em melhorias consistentes. Por exemplo, [110] aplicaram um *ensemble* por votação suave (*soft voting*) em oito modelos para o idioma hindi, obtendo ganhos significativos no F1-Score. De forma

similar, [111] utilizaram uma votação majoritária entre três modelos (como Multilingual-BERT e XLM-RoBERTa) para a tarefa de REN multilíngue em textos complexos, também obtendo melhorias consistentes em relação aos modelos individuais.

A abordagem também se mostra poderosa em domínios específicos e desafiadores. No processamento de chinês antigo, [112] testaram diferentes estratégias de votação para tarefas de etiquetagem de sequência, incluindo REN, e concluíram que os *ensembles* superaram os modelos únicos. Para o domínio de textos clínicos em chinês, [113] empregaram um método de fusão multiestratégica (uma forma de *ensemble*) sobre modelos baseados em BERT, alcançando o primeiro lugar em uma competição de REN na área médica. Um padrão recorrente nesses estudos é a aplicação de diferentes formas de **votação** como a principal técnica de combinação, validando-a como uma estratégia simples, eficaz e robusta para aprimorar o desempenho de modelos *Transformer* na tarefa de REN.

# Capítulo 3

# Trabalhos Relacionados

Este capítulo apresenta uma revisão da literatura fundamental para contextualizar a presente pesquisa. O objetivo é situar este trabalho em relação à literatura existente, identificando as lacunas de conhecimento que a dissertação se propõe a preencher. Para isso, a análise adota uma estrutura de "funil", partindo dos avanços gerais na área até o nicho específico desta investigação. A discussão está organizada em três eixos principais:

- A Seção 3.1 discute os Marcos e Tendências Atuais em REN, abordando a evolução das técnicas em nível global;
- Na Seção 3.2, a análise se concentra no domínio de aplicação, examinando o estado do REN em documentos públicos brasileiros, com foco em textos jurídicos e administrativos;
- A Seção 3.3 foca na abordagem técnica central deste trabalho, apresentando estudos que exploraram o uso de *Ensemble* de Modelos *Transformers* para REN.

Ao final, a Seção 3.4 apresenta a síntese e posicionamento deste trabalho. Nela, consolida-se a análise comparativa que evidencia a carência de estudos sobre a aplicação de *ensembles* especificamente em convênios do DOU, o que justifica a relevância científica e prática desta pesquisa.

## 3.1 Marcos e Tendências Atuais em REN

Após o estabelecimento dos modelos baseados em *Transformers* como a abordagem dominante para REN, a pesquisa de ponta na área, frequentemente avaliada no consolidado *benchmark* CoNLL-2003 [51], concentrou-se em estratégias para aprimorar ainda mais o desempenho e a robustez desses modelos. A análise de alguns trabalhos significativos

revela tendências claras que visam superar as limitações remanescentes das arquiteturas padrão. As principais são:

- **Expansão do Contexto para Nível de Documento:** Modelos de REN tradicionais analisam o texto sentença por sentença. Uma tendência recente é o desenvolvimento de arquiteturas que utilizam o contexto do documento inteiro para desambiguar entidades e melhorar a consistência das predições, como proposto em trabalhos como o FLERT [114] e o LUKE [115];
- **Otimização das Representações de Entrada (*Embeddings*):** Outra frente de pesquisa foca em aprimorar a representação numérica do texto antes de ser processado pelo *Transformer*. Isso inclui técnicas para injetar conhecimento externo ou refinar os *embeddings* para que eles já contenham informações sobre a estrutura das entidades, como explorado por [26];
- **Aumento da Robustez a Dados Ruidosos:** Sabendo que dados do mundo real raramente são perfeitos, uma importante linha de pesquisa busca desenvolver técnicas que tornem os modelos mais resilientes a ruídos, como erros de anotação ou inconsistências textuais [116]. Outra estratégia poderosa para alcançar maior robustez e aprimorar o desempenho geral é a **combinação de múltiplos modelos via *ensemble***, uma abordagem que se consolidou como prática robusta na área [14, 109] e que constitui o foco central de investigação desta dissertação.

Essas tendências demonstram que o campo de REN continua em evolução, buscando soluções cada vez mais precisas e, sobretudo, mais confiáveis para aplicações práticas.

## 3.2 REN em Documentos Públicos Brasileiros

A aplicação de REN no domínio dos documentos públicos e jurídicos no Brasil apresenta um conjunto particular de desafios. A complexidade não reside apenas nas nuances do português brasileiro, mas também na natureza dos próprios textos, caracterizados por sentenças longas, vocabulário técnico específico e estrutura frasal complexa. Esforços iniciais na área rapidamente destacaram a inadequação de ferramentas de PLN genéricas para este domínio, que falhavam em capturar as especificidades do jargão jurídico-administrativo, resultando em baixo desempenho [117, 118].

Um marco fundamental para o avanço da área foi o desenvolvimento de *corpora* anotados e de acesso público. Dentre eles, o LeNER-Br [119] se destaca como um dos mais importantes. Ao fornecer um volume significativo de textos jurídicos (decisões do Supremo Tribunal Federal, por exemplo) anotados com sete tipos de entidades (Pessoa,

Local, Organização, etc.), o LeNER-Br habilitou, pela primeira vez, o treinamento e a avaliação comparativa de modelos de aprendizado profundo para o português jurídico. Trabalhos subsequentes utilizaram este *corpus* para avaliar desde arquiteturas clássicas como BiLSTM-CRF até os modelos baseados em BERT [120].

Com a consolidação da arquitetura *Transformer*, a pesquisa se voltou para a especialização de grandes modelos de linguagem para o domínio jurídico brasileiro. Um exemplo notável é o trabalho de [121], que, ao treinar o modelo BERTimbau com um grande volume de textos em português, alcançou o estado da arte em diversas tarefas, incluindo o REN no *corpus* LeNER-Br. A eficácia de modelos BERT neste domínio foi reforçada por outras iniciativas, como a criação do LegalBERT-pt [122], que demonstraram a superioridade de modelos pré-treinados com dados de domínio específico.

Apesar desses avanços significativos na criação de recursos e na aplicação de modelos de ponta, a maior parte da pesquisa se concentrou no domínio jurídico-geral (processos, decisões, legislação) ou em nichos específicos, como a análise de editais de licitação, em que trabalhos buscaram extrair entidades como itens, valores e datas [123]. No entanto, até o início desta pesquisa, persistia uma lacuna clara na literatura: a ausência de um *corpus* e de modelos voltados à extração de um conjunto rico e detalhado de entidades do domínio específico de convênios públicos publicados no Diário Oficial da União, o que justifica a necessidade do presente trabalho.

## 3.3 Ensemble de Modelos Transformers para REN

A aplicação de *ensembles* em modelos *Transformer* para REN é uma estratégia validada na literatura recente para aprimorar a robustez e a precisão das predições. A eficácia da abordagem foi demonstrada em uma variedade de idiomas e contextos. No idioma híndi, por exemplo, [110] utilizaram *soft voting* em oito modelos, obtendo ganhos significativos. Para o chinês antigo, [112] testaram múltiplas estratégias de votação e concluíram que os *ensembles* superaram consistentemente os modelos individuais. A mesma conclusão foi alcançada em tarefas multilíngues complexas, como a do SemEval-2022, em que um *ensemble* por votação majoritária melhorou os resultados em relação a modelos individuais, como o XLM-RoBERTa [111, 124].

A estratégia também se mostrou poderosa em domínios de alta especialização. No contexto de textos clínicos em chinês, [113] empregaram uma fusão de cinco modelos baseados em BERT para reduzir o viés de um único modelo, alcançando resultados de estado da arte. Em textos matemáticos no idioma bangla, [125] utilizaram um *ensemble* de BERT que alcançou 99,76% de acurácia, superando o modelo individual em quase 2

pontos percentuais. Esses estudos consolidam o *ensemble* como uma técnica eficaz para potencializar o desempenho de modelos *Transformer* em tarefas de REN.

## 3.4 Síntese e Posicionamento deste Trabalho

A Tabela 3.1 apresenta um comparativo entre os trabalhos relacionados discutidos e a presente pesquisa. A comparação considera eixos fundamentais para a avaliação de robustez em REN: a diversidade semântica (número de classes), a escala do *corpus* (tamanho e volume de anotações) e a complexidade da arquitetura experimental (quantidade de modelos e uso de *ensembles*).

Tabela 3.1: Comparativo quantitativo e metodológico entre trabalhos relacionados e a presente dissertação.

| Referência | Domínio | Base do Corpus | Dimensão do Corpus | Entidades Anotadas | Classes | Ensemble? |
|---|---|---|---|---|---|---|
| Luz de Araujo et al. [119] | Jurídico (BR) | Docs. Legais | 318.073 tokens | N/D | 6 | Não |
| Silva et al. [120] | Licitações (BR) | Editais de Compras | 521.809 tokens | ≈ 8.300 | 8 | Não |
| Zheng et al. [112] | Histórico (CN) | Texto Antigo | 180.000 caracteres | N/D | 3 | Sim (4 modelos) |
| Sun et al. [113] | Médico (CN) | Diálogos | 2.048 sentenças | ≈ 3.700 | 18 | Sim (5 modelos) |
| Rouhizadeh et al. [111] | Geral (Multi) | Wikipedia/News | ≈ 8.800 sentenças (teste) | ≈ 250.000 | 6 | Sim (3 modelos) |
| **Este Trabalho** | **Convênios (BR)** | **DOU (1996-2022)** | **192.900 documentos** | **10.434.116** | **26** | **Sim (7 modelos)** |

Legenda: BR (Brasil), CN (China), Multi (Multilíngue). N/D: Dado não disponível explicitamente na fonte.

A análise dos dados evidencia o salto de escala proposto por esta pesquisa. Enquanto trabalhos seminais no domínio jurídico brasileiro, como [119] e [120], operam com *corpora* na casa de centenas de milhares de *tokens* e anotações na ordem de milhares, o *corpus* aqui construído ultrapassa a marca de 10,4 milhões de entidades anotadas.

Além da escala, destaca-se a granularidade da extração: são 26 classes de entidades, superando a complexidade observada em [113] (18 classes) e nas abordagens jurídicas tradicionais (6 a 8 classes). Do ponto de vista arquitetural, embora o uso de *ensembles* seja explorado na literatura internacional ([111, 112, 113]), este trabalho inova ao aplicar a combinação de sete modelos distintos baseados em *Transformers* para o contexto específico da administração pública brasileira, preenchendo a lacuna de uma abordagem que seja, simultaneamente, massiva em dados e avançada em técnica de combinação de modelos.

# Capítulo 4

# Metodologia

Neste capítulo, detalha-se a metodologia empregada na condução desta pesquisa, cujo objetivo é avaliar o uso de *ensembles* de modelos *Transformer* para a tarefa de REN em convênios públicos do DOU. O processo experimental foi organizado em quatro etapas principais, ilustradas no fluxograma da Figura 4.1 e detalhadas nas seções a seguir:

1. **Construção do *Corpus* (Seção 4.1):** Descreve-se o processo rigoroso de criação do conjunto de dados, desde o mapeamento das entidades e validação das publicações até a rotulagem, análise estatística e garantia de reprodutibilidade;
2. **Treinamento dos Modelos (Seção 4.2):** Aborda o ajuste fino (*fine-tuning*) dos modelos *Transformer* individuais que são a base para o sistema;
3. **Implementação dos *Ensembles* (Seção 4.3):** Detalha as três estratégias de combinação propostas: votação pelo melhor modelo por categoria, votação majoritária (*hard voting*) e votação ponderada;
4. **Avaliação (Seção 4.4):** Define os critérios e as métricas utilizadas para mensurar o desempenho das abordagens.

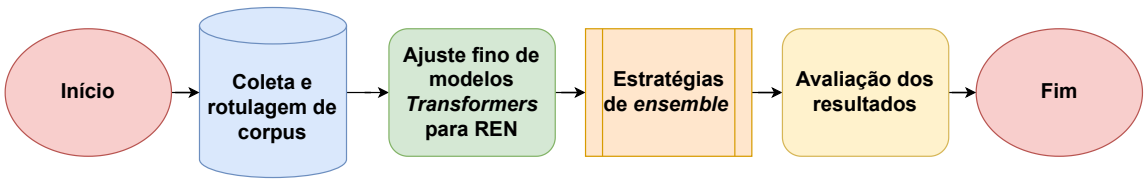



Figura 4.1: Fluxograma da Metodologia Adotada.

## 4.1 Construção do *Corpus* de Convênios do DOU

Um dos principais desafios para a aplicação de REN no domínio de convênios públicos brasileiros é a ausência de um *corpus* anotado publicamente disponível. Considerando a inviabilidade de anotar manualmente um grande volume de documentos, como os 554.101 convênios firmados entre 1996 e 2022 e identificados no Portal da Transparência [126], a primeira etapa deste trabalho consistiu no desenvolvimento de uma estratégia de anotação automática.

Este processo, realizado em colaboração com o trabalho de [4], foi dividido em quatro etapas principais, conforme ilustrado na Figura 4.2: (1) Mapeamento das Entidades Nomeadas; (2) Busca das Publicações; (3) Validação das Publicações; e (4) Rotulagem das Entidades. Cada uma dessas etapas está detalhada a seguir.

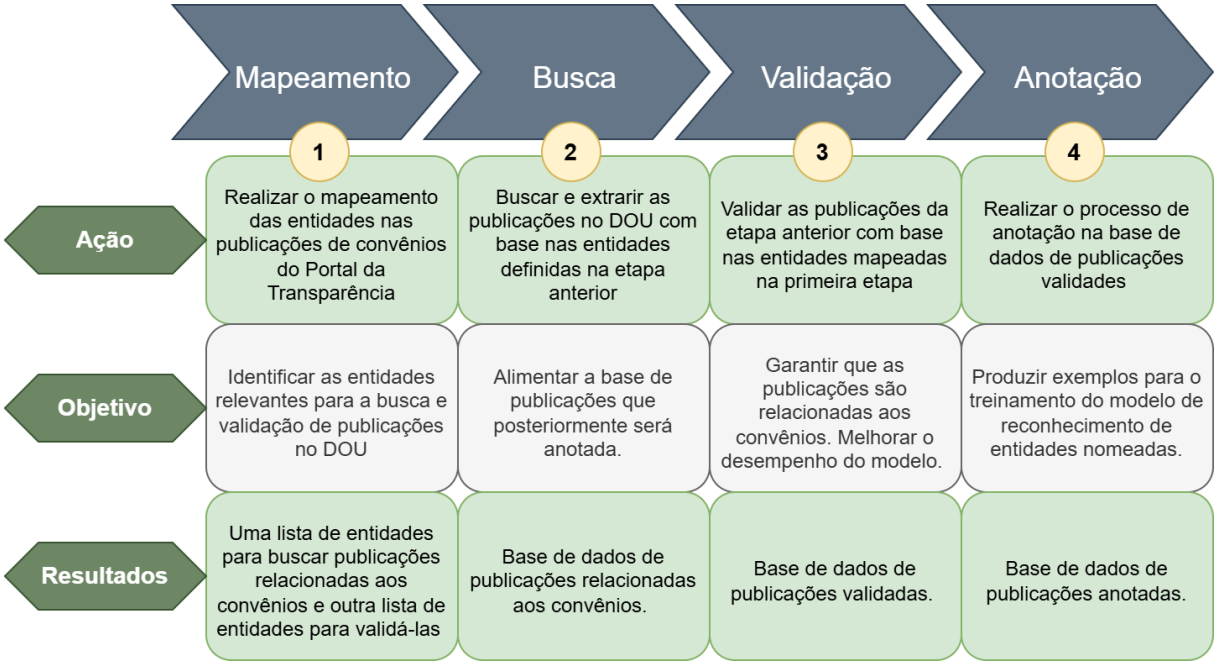



Figura 4.2: As quatro etapas do processo de anotação do *corpus*. Adaptado de: [4].

### 4.1.1 Etapa 1: Mapeamento das Entidades Nomeadas

A primeira etapa consistiu em definir o universo de entidades de interesse com base nos dados estruturados disponíveis no Portal da Transparência Brasileiro[126]. Nesta fase preliminar, ao acessar o Portal da Transparência Brasileiro, foram identificados os tipos de categorias de informações relacionadas aos convênios firmados entre 1996 e 2022, cujas definições foram validadas com o dicionário de dados oficial do portal [20].

Cabe ressaltar que, embora o mapeamento inicial tenha considerado 27 categorias presentes no Portal da Transparência Brasileiro [126], a entidade referente à "Data de

Publicação" foi posteriormente removida do conjunto final de treinamento devido à sua característica de metadado e à ausência de representatividade textual, consolidando o estudo em 26 entidades nomeadas (conforme detalhado na análise de frequências na Seção 4.1.5). A Tabela 4.1 apresenta o conjunto completo de entidades mapeadas inicialmente nesta etapa e a quantidade de anotações para cada uma dessas entidades.

Tabela 4.1: Descrição das entidades de convênios extraídas do Portal da Transparência.

| **ID da Entidade** | **Descrição** |
|---|---|
| Código Convenente | Código da entidade que recebe os recursos. |
| Código Órgão Concedente | Código do órgão que transfere os recursos. |
| Código Órgão Superior | Código do órgão superior ao qual o concedente está vinculado. |
| Código SIAFI Município | Código SIAFI que identifica o município do convenente. |
| Código UG Concedente | Código da Unidade Gestora do concedente. |
| Data Fim Vigência | Data de fim da vigência do convênio. |
| Data Início Vigência | Data de início da vigência do convênio. |
| Data Publicação | Data de publicação do convênio. |
| Data Última Liberação | Data da última liberação de recursos. |
| Nome Convenente | Nome da entidade que recebe os recursos. |
| Nome Município | Nome do município do convenente. |
| Nome Órgão Concedente | Nome do órgão que transfere os recursos. |
| Nome Órgão Superior | Nome do órgão superior ao qual o concedente está vinculado. |
| Nome UG Concedente | Nome da Unidade Gestora do concedente. |
| Número Convênio | Identificador único do convênio. |
| Número Original | Número original do convênio. |
| Número Processo do Convênio | Número do processo administrativo associado. |
| Objeto do Convênio | Descrição do objeto pactuado entre as partes. |
| Situação Convênio | Descreve o estado atual do convênio (e.g., "em execução"). |
| Tipo Convenente | Tipo da entidade convenente (e.g., "Administração Pública Municipal"). |
| Tipo Ente Convenente | Classificação do ente convenente (e.g., "MUNICIPAL"). |
| Tipo Instrumento | Tipo de instrumento legal (e.g., "CONVÊNIO"). |
| UF | Unidade Federativa do convenente. |
| Valor Contrapartida | Valor da participação financeira do convenente. |
| Valor Convênio | Valor total do convênio. |
| Valor Liberado | Valor total já liberado pelo concedente. |
| Valor Última Liberação | Valor da última liberação de recursos. |

Após a identificação das possíveis entidades, foi realizada uma análise para selecionar quais delas seriam mais eficazes para atuar como "âncoras" na busca pelas publicações correspondentes no DOU. Para isso, dois critérios foram avaliados em uma amostra de 50 convênios: unicidade e frequência. A determinação deste quantitativo considerou que, embora a redação dos atos possa apresentar variabilidade sintática, a presença dos dados cadastrais essenciais obedece a requisitos legais de publicação. Dessa forma, a inspeção manual de 50 documentos mostrou-se suficiente para identificar quais atributos aparecem

de forma consistente e única, permitindo sua validação como chaves de busca eficientes, atingindo-se rapidamente um consenso sobre quais metadados utilizar.

### Análise de Unicidade

A unicidade mede a capacidade de uma entidade de identificar publicações pertencentes a um único convênio, minimizando a recuperação de documentos irrelevantes. Para avaliar este critério, foi definida a métrica "Porcentagem de Relevância" ($PR$), calculada para cada tipo de entidade $E_i$ em uma amostra de $n$ convênios, conforme apresentado na Equação 4.1.

$$PR(E_i) = \frac{\sum_{j=1}^{n} |R(E_i, C_j)|}{\sum_{j=1}^{n} |P(E_i, C_j)|} \times 100 \tag{4.1}$$

onde $|R(E_i, C_j)|$ refere-se ao número de publicações relevantes obtidas ao buscar pela entidade $E_i$ do convênio $C_j$, e $|P(E_i, C_j)|$ representa o número total de publicações retornadas por essa pesquisa. Foi utilizada a amostra de $n = 50$ convênios para esta análise.

### Análise de Frequência

A frequência mede a probabilidade de uma entidade aparecer nas publicações textuais de um determinado convênio. Uma entidade, mesmo que única, tem pouca utilidade se não for publicada de forma consistente. A frequência $F$ de uma entidade $E_i$ em um dado convênio $C_j$ foi calculada conforme a Equação 4.2.

$$F(E_i, C_j) = \frac{|P_{ocorre}(E_i, C_j)|}{|P_{total}(C_j)|} \tag{4.2}$$

onde $|P_{ocorre}(E_i, C_j)|$ é o número de publicações do convênio $C_j$ em que a entidade $E_i$ está presente, e $|P_{total}(C_j)|$ é o número total de publicações relacionadas ao convênio $C_j$.

A análise de unicidade e frequência, juntamente com a observação de variações de formato (por exemplo, CNPJ com e sem pontuação), foi essencial para selecionar as entidades-chave que orientaram a próxima etapa de busca por publicações.

### Seleção e Análise de Entidades-Chave

Para selecionar os tipos de entidades mais eficazes para as etapas de busca e validação, foi conduzida uma análise empírica sobre a amostra de $n = 50$ convênios, avaliando-se os critérios de unicidade e de frequência.

Os limiares para a classificação de cada entidade como "Alto", "Médio" ou "Baixo" foram definidos de forma heurística, com base na observação da distribuição dos valores na

amostra. Buscou-se criar faixas que separassem os grupos de entidades com comportamentos claramente distintos. Para a unicidade, os limiares foram: Baixo ($PR < 30\%$), Médio ($30\% \leq PR \leq 80\%$) e Alto ($PR > 80\%$). Para a frequência, os limiares foram: Baixo ($F < 0.2$), Médio ($0.2 \leq F \leq 0.6$) e Alto ($F > 0.6$). A Tabela 4.2 resume a classificação resultante dessa análise.

Tabela 4.2: Resultado do mapeamento das entidades. Fonte: [4].

| **Entidade** | **Unicidade** | **Frequência** |
|---|---|---|
| Número Convênio | **Alto** | **Alto** |
| UF | **Baixo** | **Alto** |
| Código SIAFI Município | **Baixo** | **Baixo** |
| Nome Município | **Baixo** | **Médio** |
| Situação Convênio | **Baixo** | **Baixo** |
| Número Original | **Médio** | **Baixo** |
| Número Processo do Convênio | **Alto** | **Médio** |
| Objeto do Convênio | **Alto** | **Médio** |
| Código Órgão Superior | **Baixo** | **Baixo** |
| Nome Órgão Superior | **Baixo** | **Alto** |
| Código Órgão Concedente | **Baixo** | **Baixo** |
| Nome Órgão Concedente | **Baixo** | **Médio** |
| Código UG Concedente | **Baixo** | **Médio** |
| Nome UG Concedente | **Baixo** | **Baixo** |
| Código Convenente | **Médio** | **Médio** |
| Tipo Convenente | **Baixo** | **Baixo** |
| Nome Convenente | **Baixo** | **Médio** |
| Tipo Ente Convenente | **Baixo** | **Alto** |
| Tipo Instrumento | **Baixo** | **Baixo** |
| Valor Convênio | **Baixo** | **Alto** |
| Valor Liberado | **Baixo** | **Médio** |
| Data Publicação | **Baixo** | **Baixo** |
| Data Início Vigência | **Baixo** | **Alto** |
| Data Fim Vigência | **Baixo** | **Médio** |
| Valor Contrapartida | **Baixo** | **Médio** |
| Data Última Liberação | **Baixo** | **Baixo** |
| Valor Última Liberação | **Baixo** | **Alto** |

Com base nesta análise, as entidades foram selecionadas para papéis específicos. Para

a busca de publicações, foram escolhidas as cinco entidades que apresentaram nível de unicidade Médio ou Alto: 'Número Convênio', 'Número Original', 'Número Processo do Convênio', 'Objeto do Convênio' e 'Código Convenente'. Para a validação, foram selecionadas treze entidades com nível de frequência Médio ou Alto que não foram utilizadas na busca. As nove entidades restantes foram excluídas dessas etapas, mas mantidas para a rotulagem final.

### 4.1.2 Etapa 2: Busca das Publicações

Uma vez definidas as entidades-âncora, a segunda etapa da metodologia concentrou-se em localizar e extrair, de forma sistemática, as publicações textuais de cada convênio a partir do portal do DOU [9]. Para executar essa tarefa em larga escala, foi empregada a técnica de raspagem de dados (*web scraping*) [127], que permite a extração automatizada de conteúdo de páginas web.

A estratégia de busca fundamentou-se no uso das cinco entidades de alta e média unicidade (e.g., 'Número Convênio', 'Número Processo do Convênio') como termos de pesquisa. Para assegurar a consistência e a reprodutibilidade dos resultados, todas as buscas foram parametrizadas com um conjunto fixo de filtros, detalhados na Tabela 4.3. A aplicação desses filtros, como a restrição à "Seção 3" do diário e a busca por "Resultado exato", foi crucial para refinar o universo de resultados e focar apenas nos documentos de maior relevância potencial.

O processo técnico de extração foi implementado por meio de um algoritmo iterativo, cujo fluxo operacional detalhado é apresentado na Figura 4.3. Em essência, para cada convênio encontrado no Portal da Transparência, o sistema itera sobre suas entidades-âncora, construindo URLs de requisição HTTP para consultar o portal do DOU. A partir da página de resultados, são extraídos os *links* para cada publicação individual. Em seguida, o sistema acessa cada um desses *links* para extrair e armazenar o texto completo, mantendo sempre a associação com o convênio de origem. O algoritmo também inclui uma verificação para evitar o armazenamento de publicações duplicadas. Ao final desta etapa, obteve-se um grande conjunto de publicações candidatas, prontas para a fase de validação.

### 4.1.3 Etapa 3: Validação das Publicações

A etapa de busca, embora otimizada, inevitavelmente recupera publicações que não pertencem ao convênio de interesse (falsos positivos), dado que as entidades-âncora podem não ser absolutamente únicas no universo do DOU. Para mitigar esse ruído, a terceira etapa da metodologia consistiu na aplicação de um filtro de validação heurística. O prin-

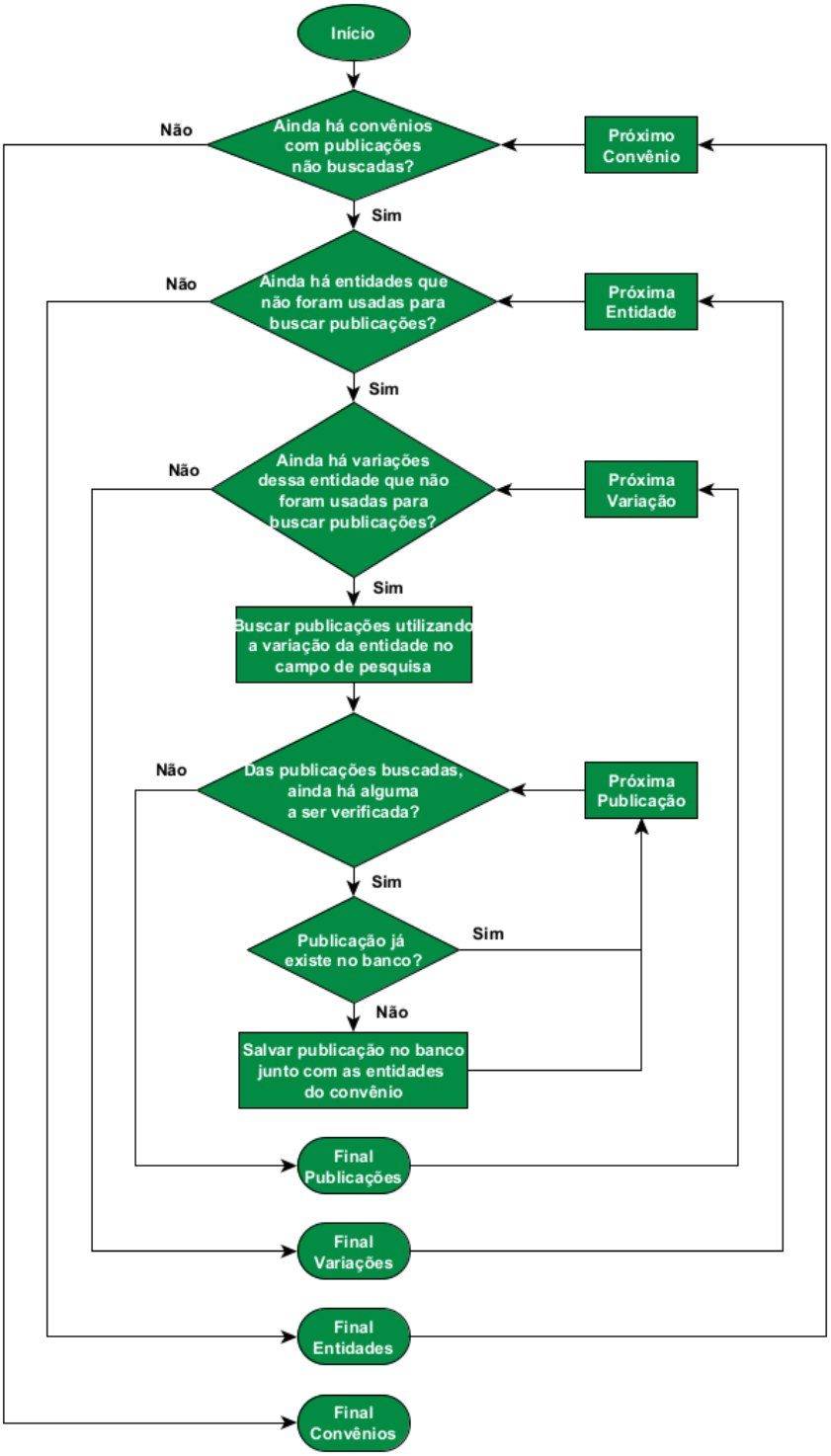



Figura 4.3: Fluxograma do algoritmo de busca e extração de publicações. O processo itera sobre convênios, entidades-âncora e suas variações para garantir uma busca abrangente. Adaptado de: [4].

cípio do método consistiu em utilizar o conjunto secundário de 13 entidades de alta e média frequência, denominadas "entidades de validação", para confirmar a pertinência de

Tabela 4.3: Filtros de pesquisa disponíveis no portal do Diário Oficial da União.

| **Filtro** | **Valor Utilizado** | **Justificativa** |
|---|---|---|
| Tipo de pesquisa | Resultado exato | Retorna resultados mais precisos, evitando publicações parcialmente relacionadas. |
| Forma de pesquisa | Pesquisa ato a ato | Exibe as publicações como páginas web individuais, possibilitando a aplicação do *web scraping*. |
| Onde pesquisar | Tudo | Garante que a entidade seja buscada tanto no título quanto no corpo do texto da publicação. |
| Ordenação | Por data | Parâmetro padrão sem impacto direto na coleta, mas mantido para consistência. |
| Data | Personalizado: 01/01/2018 até a data da pesquisa | A forma de pesquisa "ato a ato" está disponível apenas para publicações a partir de 2018. |
| Jornal | Seção 3 | Seção onde são publicados os extratos de contratos, convênios e instrumentos similares. |

cada documento. A premissa é que a coocorrência de múltiplas informações de um mesmo convênio em um único texto é um forte indicativo de sua relevância.

O mecanismo de validação foi implementado da seguinte forma: para cada publicação candidata, o algoritmo verifica a presença das entidades de validação. Uma publicação é considerada válida somente se é encontrado, em seu conteúdo, um limiar mínimo de três entidades de validação distintas. Documentos que não atingem esse limiar são descartados do conjunto de dados. Conforme detalhado em [4], este limiar foi definido empiricamente após testes em uma amostra de 50 convênios, buscando o melhor equilíbrio entre remover publicações irrelevantes (aumentar a precisão) e manter as corretas (não prejudicar a cobertura).

O fluxograma que detalha a implementação deste algoritmo de validação é apresentado na Figura 4.4. Ele ilustra a lógica iterativa e a regra de decisão, incluindo uma otimização de desempenho: a verificação de um documento é interrompida e ele é aprovado assim que o limiar de três entidades é atingido (*early exit*), sem a necessidade de processar as entidades de validação restantes para aquele texto.

### 4.1.4 Etapa 4: Rotulagem das Entidades nas Publicações

A etapa final do processo de construção do *corpus* foi a rotulagem, na qual as informações estruturadas do Portal da Transparência foram utilizadas para anotar as entidades correspondentes nos textos validados do DOU. O objetivo desta fase foi transformar as publicações em formato textual em um conjunto de dados estruturado, pronto para o treinamento dos modelos.

Este processo foi implementado utilizando a ferramenta `PhraseMatcher` da biblioteca *spaCy* [40], que é otimizada para encontrar sequências de texto de um dicionário de forma eficiente. O algoritmo de rotulagem foi executado da seguinte maneira: para cada

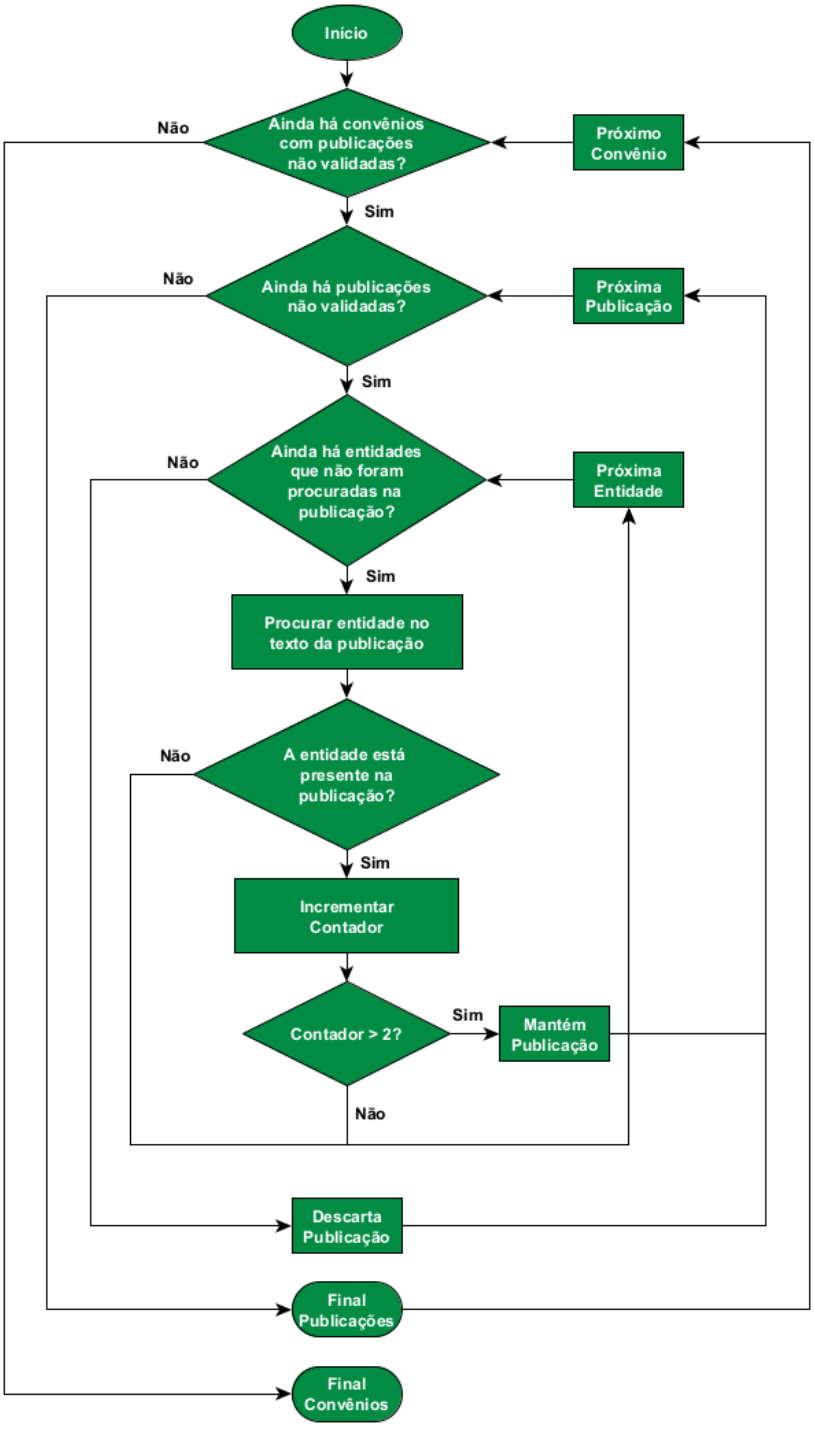



Figura 4.4: Fluxograma do algoritmo de validação heurística. O processo descarta ou mantém uma publicação com base na contagem de "entidades de validação" presentes em seu texto. Adaptado de: [4].

convênio, uma instância do *PhraseMatcher* foi criada e alimentada com os valores de suas possíveis entidades (e.g., "Nome do Convenente", "Valor do Convênio", "Número do Processo", etc.), extraídos dos dados estruturados do Portal da Transparência. Cada valor foi associado ao respectivo rótulo de entidade. Em seguida, o *PhraseMatcher* foi executado sobre cada publicação validada do convênio, identificando todas as ocorrências de correspondência exata e retornando seus índices de início e fim no texto.

O resultado deste processo, para cada documento, foi uma lista de anotações no formato (`rótulo_da_entidade`, `índice_inicial`, `índice_final`). A Figura 4.5 apresenta uma visualização de como um texto fica após a aplicação dessas anotações.

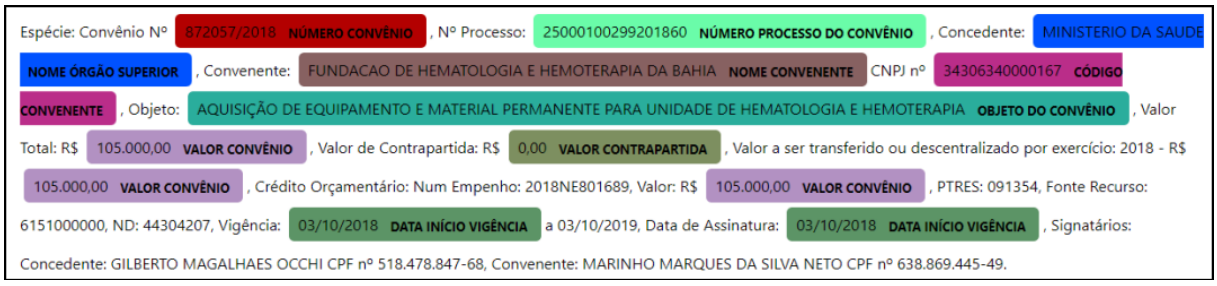



Figura 4.5: Visualização de uma publicação anotada, onde cada cor representa um tipo de entidade. Fonte: [4].

Para o treinamento com o *framework spaCy*, esses dados foram estruturados no formato exigido pela ferramenta. Cada exemplo de treinamento consiste em um texto e um dicionário associado, contendo uma lista de tuplas, em que cada tupla especifica os índices de início e fim (em caracteres) e o rótulo de uma entidade. Este formato, que utiliza índices de caracteres, é robusto a diferentes estratégias de tokenização [40]. Um exemplo simplificado da estrutura de dados de uma sentença é apresentado na Figura 4.6.

```
# Consiste em uma tupla: (texto, dicionário_de_anotações)
texto = "Eutino Júnior nasceu em Corrente, no Piauí."

training_example = (
    texto,
    {
        "entities": [
            (0, 14, "PESSOA"),
            (25, 33, "LOCAL"),
            (38, 43, "LOCAL"),
        ]
    },
)
# As entidades são uma lista de tuplas: (índice_inicial, índice_final, rótulo)
```

Figura 4.6: Exemplo da estrutura de dados de treinamento no formato do spaCy.

Ao final deste processo de quatro etapas, foi consolidado o *corpus* final utilizado nos experimentos de treinamento e avaliações.

### 4.1.5 Análise Descritiva e Estatística do *Corpus* Final

Ao final do processo de quatro etapas, foi consolidado o *corpus* final. No total, foram coletadas e anotadas 192.900 publicações textuais, recuperadas a partir de um universo de busca de 554.101 convênios firmados entre 1996 e 2022. Essas publicações correspondem a 71.287 convênios distintos, resultando em uma média de 2,7 publicações por convênio encontrado. A Tabela 4.4 apresenta a distribuição quantitativa das anotações. O número de anotações refere-se à quantidade total de vezes que uma entidade foi rotulada em todo o *corpus*.

A análise da distribuição revela um severo desequilíbrio entre as classes. Entidades como 'Tipo Ente Convenente' e 'Data Início Vigência' somam, juntas, mais de 4,4 milhões de anotações (aproximadamente 51% do total), enquanto 12 das 27 entidades possuem menos de 1% de representatividade cada uma. Este desbalanceamento é uma característica intrínseca do domínio e impõe um desafio significativo para o treinamento de modelos de REN.

Tabela 4.4: Distribuição final de anotações por entidade no *corpus*.

| **Entidade** | **Qtde. de Anotações** | **Percentual (%)** |
|---|---:|---:|
| Tipo Ente Convenente | 2.274.833 | 21,80% |
| Data Início Vigência | 2.190.278 | 20,99% |
| Nome Órgão Superior | 1.495.540 | 14,33% |
| UF | 1.361.105 | 13,04% |
| Valor Última Liberação | 1.275.212 | 12,22% |
| Data Fim Vigência | 508.623 | 4,87% |
| Código UG Concedente | 301.681 | 2,89% |
| Valor Convênio | 171.283 | 1,64% |
| Nome Convenente | 149.594 | 1,43% |
| Valor Contrapartida | 137.369 | 1,32% |
| Objeto do Convênio | 114.439 | 1,10% |
| Código Convenente | 102.837 | 0,99% |
| Número Convênio | 92.891 | 0,89% |
| Nome Município | 68.081 | 0,65% |
| Nome Órgão Concedente | 66.591 | 0,64% |
| Número Processo do Convênio | 38.371 | 0,37% |
| Número Original | 25.771 | 0,25% |
| Código SIAFI Município | 23.166 | 0,22% |
| Código Órgão Concedente | 17.774 | 0,17% |

Tabela 4.4: Distribuição final de anotações por entidade no *corpus*.

| **Entidade** | **Qtde. de Anotações** | **Percentual (%)** |
|---|---:|---:|
| Data Última Liberação | 8.763 | 0,08% |
| Situação Convênio | 4.004 | 0,04% |
| Tipo Instrumento | 3.059 | 0,03% |
| Nome UG Concedente | 1.383 | 0,01% |
| Código Órgão Superior | 1.114 | 0,01% |
| Tipo Convenente | 191 | $< 0{,}01\%$ |
| Valor Liberado | 162 | $< 0{,}01\%$ |
| Data Publicação | 1 | $< 0{,}01\%$ |
| **Total** | **10.434.116** | **100,00%** |

**Limitações danotação Automática e Estratégia A de Mitigação**

É fundamental reconhecer uma limitação inerente ao método de anotação automática por correspondência exata, detalhado na Etapa 4 (Seção 4.1.4). A ferramenta `PhraseMatcher` operou buscando a ocorrência literal dos textos extraídos do Portal da Transparência. Conforme detalhado no trabalho de [4], essa busca foi aprimorada para reconhecer um conjunto de variações textuais comuns e predefinidas em certas entidades, como os diferentes formatos de um CNPJ. Apesar dessa melhoria, a abordagem permaneceu sensível a qualquer variação textual não prevista — como abreviações não mapeadas, erros de digitação ou outras diferenças mínimas de formatação — que, consequentemente, ainda resultou em uma limitação de anotação.

Este fenômeno gera um gabarito com "falsos negativos", ou seja, um *corpus* em que a anotação intencionalmente prioriza a precisão em detrimento da revocação. Em outras palavras, as anotações feitas são de alta confiança (alta precisão), mas nem todas as entidades que deveriam ser anotadas, de fato, o são (baixa revocação).

A decisão por essa abordagem foi uma escolha metodológica pragmática, impulsionada pela escala monumental do problema. A anotação manual das publicações correspondentes ao universo de 554.101 convênios, que possui cada uma múltiplos documentos vinculados, seria uma tarefa inexequível no escopo deste projeto. A estratégia adotada, portanto, foi substituir a perfeição em nível de documento pela força estatística em nível de *corpus*. A hipótese central foi que a exposição dos modelos a uma escala massiva de dados compensaria o ruído e a incompletude de anotações individuais. Esta premissa está alinhada aos princípios da Supervisão a Distância (*Distant Supervision*) [128], em que uma base de conhecimento externa (neste caso, o Portal da Transparência) é utilizada

para gerar automaticamente dados de treinamento heurísticos para textos não rotulados (as publicações do DOU).

A principal estratégia de mitigação da limitação do gabarito foi, portanto, o próprio volume do *corpus*. Mesmo que uma entidade seja omitida em um documento, a alta probabilidade de que ela apareça corretamente anotada em milhares de outros documentos permite que os modelos *Transformer* aprendam os padrões contextuais e semânticos que definem essa entidade. A eficácia desta abordagem é comprovada empiricamente nos resultados, onde a análise qualitativa (Seção 5.3) demonstra que os modelos treinados foram capazes de generalizar o aprendizado para além da correspondência exata, identificando corretamente entidades que estavam ausentes no gabarito original.

### Análise de Entidades de Baixa Quantidade de Anotações

A análise qualitativa de entidades com pouquíssimas anotações revelou *insights* importantes sobre o processo de REN. As três classes de entidades com a menor quantidade de dados anotados são apresentadas nos itens a seguir.

- **Data Publicação (1 anotação):** A investigação demonstrou que a data de publicação de um ato no DOU é um metadado da publicação e não está presente no corpo do texto. A única anotação registrada provavelmente ocorreu por coincidência de formato com outra data. Devido à sua irrelevância e à ausência de exemplos, esta entidade foi **removida do conjunto de dados** antes do treinamento dos modelos, resultando em 26 tipos de entidades nomeadas.
- **Valor Liberado (162 anotações) e Tipo Convenente (191 anotações):** A baixa frequência sugere que tais informações, embora presentes nos dados estruturados do Portal da Transparência, não são comumente detalhadas ou especificadas nos textos das publicações oficiais.

### Implicações para o Treinamento dos Modelos

A caracterização do *corpus* é fundamental para contextualizar os resultados apresentados no Capítulo 5. O grande volume de dados e a alta frequência de certas entidades oferecem uma base robusta para o treinamento dos modelos. Contudo, o severo desequilíbrio de classes é o principal desafio a ser superado. Modelos de REN podem desenvolver um viés em favor das classes majoritárias, resultando em desempenho inferior para as entidades minoritárias. A performance dos modelos diante deste cenário e como as técnicas de *ensemble* tentam mitigar esse problema é o foco das seções e capítulos subsequentes.

### 4.1.6 Disponibilidade do *Corpus* e Reprodutibilidade

Visando assegurar a transparência científica e viabilizar a reprodutibilidade dos experimentos apresentados, o *corpus* consolidado de convênios foi publicado em repositório de dados abertos.

O conjunto de dados, composto por 192.900 publicações de documentos de convênios extraídas do DOU, abrange um total de 10.434.116 entidades anotadas distribuídas em 26 classes. O material encontra-se disponível no Zenodo sob a licença *Creative Commons Attribution 4.0 International*[1].

Para garantir a consistência experimental em trabalhos futuros, os dados foram segregados fisicamente e disponibilizados na mesma partição utilizada nesta pesquisa:

- **Conjunto de Treino (70%):** Utilizado para o ajuste dos modelos base;
- **Conjunto de Teste (30%):** Preservado para a avaliação final dos sistemas e dos *ensembles*.

O acesso aos dados e aos metadados completos pode ser realizado através do identificador de objeto digital (DOI)[2].

## 4.2 Treinamento dos Modelos Base para o REN

Com o *corpus* devidamente construído e anotado, a etapa seguinte da metodologia consiste no ajuste fino (*fine-tuning*) de modelos pré-treinados para a tarefa de REN no domínio específico de convênios. Este processo, fundamentado no paradigma de Transferência de Aprendizagem apresentado na Seção 2.5.4, visa especializar o conhecimento linguístico de grandes modelos de linguagem para o contexto específico deste trabalho.

### 4.2.1 Modelos *Transformer* Utilizados

Foram selecionados sete modelos pré-treinados baseados na arquitetura *Transformer*, buscando cobrir uma diversidade de abordagens, como modelos multilíngues, modelos focados no português do Brasil e modelos especializados no domínio jurídico. A Tabela 4.5 detalha os sete modelos base utilizados nos experimentos.

[1] A licença CC BY 4.0 permite que terceiros distribuam, remixem, adaptem e criem a partir do material, mesmo para fins comerciais, desde que atribuam o devido crédito ao autor original. Fonte: `https://creativecommons.org/licenses/by/4.0/`.

[2] Repositório: Zenodo (*Corpus* Anotado de Convênios do DOU). DOI: `https://doi.org/10.5281/zenodo.17772231`

Tabela 4.5: Modelos *Transformer* base utilizados para o ajuste fino.

| ID | Modelo | Referência |
|---|---|---|
| M1 | `Albert-base-v2` | [129] |
| M2 | `bert-base-cased` | [88] |
| M3 | `Electra-base-discriminator` | [130] |
| M4 | `bart-base` | [91] |
| M5 | `legal-bert-base-cased-ptbr` | [131] |
| M6 | `bert-base-portuguese-cased` | [121] |
| M7 | `xlm-roberta-base` | [132] |

### 4.2.2 Processo de Ajuste Fino dos Modelos Transformers

O procedimento adotado seguiu o paradigma de Transferência de Aprendizagem, partindo dos pesos dos modelos pré-treinados (descritos na Seção 4.2.1) para especializar o conhecimento linguístico no domínio dos convênios. Conforme fundamentado na Seção 2.5.4, que apresenta a arquitetura *Transformers* e os métodos de REN nela baseados, a abordagem adotada empregou modelos desta classe como codificadores contextuais, alimentando o componente de Reconhecimento de Entidades Nomeadas gerenciado pelo *framework* spaCy[40].

No que tange aos dados para o treinamento, adotou-se a anotação baseada em *spans* (intervalos), formato nativo da ferramenta. Este formato especifica diretamente os índices de início e fim da entidade em nível de caracteres. O ciclo de ajuste fino foi conduzido de forma iterativa e supervisionada. Para garantir a comparabilidade entre os diferentes experimentos, não foi realizada busca individual de hiperparâmetros para cada modelo pré-treinado. Em vez disso, utilizou-se um conjunto robusto de configurações mantidas constantes para todos os modelos, conforme justificado na Seção 4.2.3 e detalhado na Tabela 4.6.

A otimização dos pesos foi realizada pelo algoritmo Adam [133], configurado com uma estratégia de *warmup* linear na taxa de aprendizado. Esta técnica estabiliza os gradientes nas etapas iniciais do ajuste fino, evitando divergências bruscas nos pesos pré-treinados.

Um aspecto técnico relevante é a utilização de acumulação de gradiente (*gradient accumulation*). Para contornar limitações de memória da GPU e, ao mesmo tempo, garantir uma estimativa de erro estável, o tamanho efetivo do lote (*batch size*) é fixado em 96. Isso foi obtido ao processar-se 3 sublotes de 32 amostras antes de cada atualização dos pesos. Além disso, para mitigar o sobreajuste (*overfitting*), aplicou-se a estratégia de parada antecipada (*early stopping*) com uma "paciência" de 1.600 passos. Isso significa que o treinamento é interrompido caso a métrica de avaliação não melhore após 8 ciclos consecutivos de validação.

Conforme ilustrado no fluxo da Figura 4.7, a cada 200 passos, o modelo foi avaliado no conjunto de validação, salvando-se o estado (*checkpoint*) que maximizou a métrica *F1-Score*.

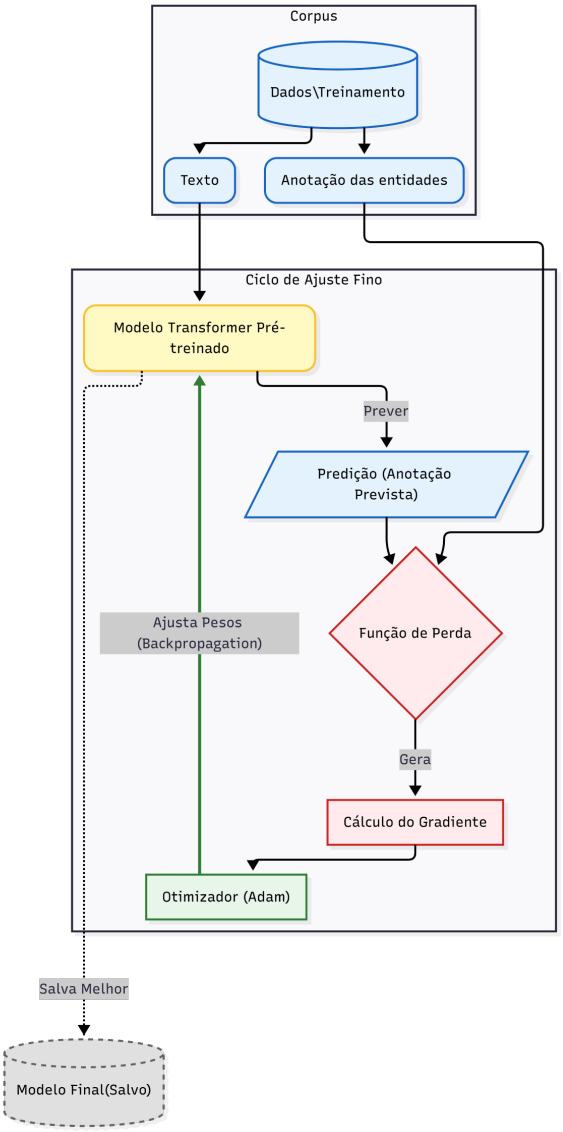



Figura 4.7: Fluxograma do processo de ajuste fino supervisionado.

Em síntese, conforme ilustrado na Figura 4.7, o ciclo de ajuste fino inicia-se com a segregação dos dados de treinamento em duas vertentes: o texto bruto, que alimenta o modelo, e as anotações de referência. O modelo processa o texto para gerar uma predição, que é então confrontada às anotações reais na função de perda. O gradiente dessa perda é calculado e utilizado pelo otimizador para ajustar os pesos internos da rede, executando a etapa de *backpropagation* [134]. Este ciclo se repete por múltiplas épocas e, conforme indicado pela conexão pontilhada no diagrama, o estado do modelo que apresenta o melhor

desempenho na validação é salvo como o resultado final, garantindo a preservação da melhor capacidade de generalização alcançada [40].

### 4.2.3 Ambiente Experimental e Hiperparâmetros

Para garantir a reprodutibilidade dos experimentos, todos os sete modelos foram treinados e avaliados sob as mesmas condições. O treinamento foi conduzido utilizando a biblioteca **spaCy** (versão 3.7.4) [40], com o alocador de memória da GPU configurado para `pytorch`. O ambiente experimental contou com as seguintes especificações de hardware:

- **GPU:** uma NVIDIA Tesla V100S com 32 GB de VRAM;
- **CPU:** Intel Xeon Gold 5220R @ 2.20GHz (utilizando 48 CPUs);
- **Memória RAM:** 376 GB.

Para a definição dos hiperparâmetros, utilizou-se o sistema de recomendação nativo da biblioteca **spaCy**, acessado pelo comando `init config`. A escolha por manter as configurações sugeridas por esta ferramenta justifica-se por dois motivos fundamentais:

1. **Base em Melhores Práticas:** Conforme a documentação oficial da ferramenta, as recomendações geradas pelo `init config` baseiam-se em extensos experimentos e "melhores práticas" gerais identificadas pelos desenvolvedores para equilibrar eficiência e precisão [40], fornecendo padrões robustos para tarefas de PLN;
2. **Isolamento de Variáveis para Comparação:** Como o objetivo central do experimento é comparar o desempenho de diferentes arquiteturas de modelos de linguagem (e, posteriormente, seus *ensembles*), foi imperativo manter um conjunto fixo de hiperparâmetros. Isso garante que as variações de desempenho observadas na Tabela de Resultados sejam atribuídas estritamente às capacidades das arquiteturas dos modelos pré-treinados, e não a variações estocásticas ou manuais nos parâmetros de ajuste fino.

Os principais hiperparâmetros resultantes dessa configuração, mantidos constantes em todos os experimentos, estão detalhados na Tabela 4.6. A avaliação final de cada modelo foi realizada no conjunto de teste. O desempenho foi medido com base nas métricas de Precisão, Revocação e *F1-Score*, sob o critério de avaliação exata, conforme detalhado na Seção 2.4.1.

Tabela 4.6: Hiperparâmetros utilizados no processo de ajuste fino.

| **Hiperparâmetro** | **Valor** |
|---|---|
| Otimizador | Adam [133] |
| Taxa de Aprendizagem (*Learning Rate*) | Inicial de 5e-5 com *warmup* linear |
| Tamanho Efetivo do Lote (*Batch Size*) | 96 (32 *batch size* × 3 *accumulate gradient*) |
| Máximo de Passos (*Max Steps*) | 20.000 |
| Frequência de Avaliação | A cada 200 passos |
| Dropout | 0.1 |
| Regularização L2 | 0.01 |
| Gradiente Clip | 1.0 |
| Paciência (*Patience*) | 1.600 passos |

## 4.3 Implementação das Estratégias de *Ensemble*

Após a avaliação dos modelos individuais, a etapa final da metodologia consistiu na implementação de estratégias de *ensemble* para combinar suas previsões. A motivação para esta abordagem foi dupla. Primeiramente, como demonstrado na revisão da literatura (Capítulo 3), a combinação de múltiplos modelos *Transformer* é uma técnica consolidada para aprimorar o desempenho em tarefas de REN. Em segundo lugar, do ponto de vista prático, a aplicação de *ensembles* aos modelos já treinados representou uma alternativa mais eficiente em termos de custo computacional do que a reexecução de um ciclo completo de ajuste de hiperparâmetros para cada um dos sete modelos.

Dessa forma, foram desenvolvidas e avaliadas três estratégias distintas, todas baseadas nos princípios de votação discutidos na Seção 2.6. A Tabela 4.7 resume as abordagens implementadas, detalhadas nas subseções a seguir.

Tabela 4.7: Resumo das estratégias de *ensemble* implementadas.

| **ID** | **Tipo de Ensemble** | **Regra de Decisão Principal** |
|---|---|---|
| E1 | Votação do Melhor Modelo | Maior *F1-Score* por classe |
| E2 | Votação por Maioria | Metade mais um dos votos |
| E3 | Votação Ponderada | Média do *F1-Score* dos votantes $> 0.8$ |

### 4.3.1 Estratégia 1: Votação pelo Melhor Modelo por Categoria

A primeira e mais original estratégia de *ensemble* implementada foi uma abordagem especialista, aqui denominada Votação pelo Melhor Modelo por Categoria. A premissa desta técnica é que, devido às suas arquiteturas e aos dados de pré-treinamento distintos, alguns modelos podem se tornar mais proficientes em reconhecer certas categorias de entidades do que outros. Em vez de buscar um consenso geral, este método delega a decisão final

ao modelo que apresentou o melhor desempenho para cada classe de entidade específica. O método foi implementado em duas fases:

1. **Fase de Calibração (Pré-requisito):** Primeiramente, foi realizada uma análise de desempenho detalhada para cada um dos sete modelos base ($M_1, \ldots, M_7$) no conjunto de teste. Para cada uma das categorias de entidade $c_k$ (onde $k = 1, \ldots, 26$), o modelo com o maior *F1-Score* foi identificado e designado como o "modelo especialista" para aquela categoria. Formalmente, seja $F(M_i, c_k)$ o *F1-Score* do modelo $M_i$ para a classe $c_k$. O modelo especialista para a classe $c_k$, denotado por $M^*_{c_k}$, é determinado pela Equação 4.3.

$$M^*_{c_k} = \text{argmax}_{M_i} F(M_i, c_k) \tag{4.3}$$

   O resultado desta fase foi um mapa que associa cada classe de entidade ao seu respectivo modelo de melhor desempenho.

2. **Fase de Predição e Filtragem:** Em seguida, para um novo documento de entrada, todos os sete modelos geraram suas respectivas listas de predições de forma independente. Seja $E_i$ o conjunto de entidades previsto pelo modelo $M_i$. O algoritmo de *ensemble* então construiu o conjunto de predições finais, $E_{final}$, aplicando uma regra de filtragem: uma predição de entidade $e$ foi mantida na saída final **somente se** ela foi gerada pelo modelo previamente designado como o especialista para o seu respectivo rótulo. Todas as predições feitas por modelos não especialistas para uma dada categoria foram descartadas. Formalmente, o conjunto final é a união das predições de cada especialista, restrito à sua classe de especialidade, como descrito na Equação 4.4.

$$E_{final} = \bigcup_{k=1}^{27} \{e \in E_{M^*_{c_k}} \mid \text{rótulo}(e) = c_k\} \tag{4.4}$$

Dessa forma, a predição final do *ensemble* E1 é uma "colcha de retalhos" composta pelas melhores previsões de cada especialista em sua área de proficiência. O objetivo desta técnica é maximizar o desempenho por classe, confiando na decisão do modelo que demonstrou maior aptidão para cada tipo de entidade.

### 4.3.2 Estratégia 2: Votação por Maioria (*Hard Voting*)

A Votação por Maioria, também conhecida como *Hard Voting*, é a estratégia de *ensemble* mais direta e foi a segunda abordagem implementada. O princípio fundamental é que uma

predição só é considerada confiável se a maioria absoluta dos modelos base concordar com ela, tanto em seus limites (posição) quanto em seu tipo (rótulo) [103].

O algoritmo implementado opera da seguinte forma: para um dado texto de entrada, são, primeiramente, coletadas as listas de entidades previstas por cada um dos $n = 7$ modelos base. Seja $E_i$ o conjunto de entidades previstas pelo modelo $M_i$, em que cada entidade $e$ é uma tupla única definida por (`índice_inicial, índice_final, rótulo`). O sistema então calcula o número de votos $k(e)$ para cada entidade única $e$ presente no conjunto total de previsões de todos os modelos. A contagem de votos é formalizada pela Equação 4.5, que utiliza a função indicadora ($\mathbf{1}$) para somar 1 a cada modelo que previu a entidade $e$.

$$k(e) = \sum_{i=1}^{n} \mathbf{1}_{e \in E_i} \tag{4.5}$$

A regra de decisão final estipula que uma entidade $e$ só é mantida na saída do *ensemble* se o número de votos $k(e)$ for estritamente maior que a metade do número total de modelos. Formalmente, a entidade é aceita se:

$$k(e) > \frac{n}{2} \tag{4.6}$$

Considerando $n = 7$ modelos, o limiar para aceitação de uma entidade é de, no mínimo, 4 votos.

O objetivo desta abordagem conservadora é aumentar a Precisão do sistema final, filtrando predições falsas ou de baixa confiança produzidas por apenas uma minoria dos modelos. Embora seja robusta, essa estratégia pode, por outro lado, reduzir a Revocação ao descartar entidades válidas para as quais não houve consenso majoritário.

### 4.3.3 Estratégia 3: Votação Ponderada por Desempenho

A terceira estratégia implementada foi uma forma de votação ponderada que utiliza o desempenho histórico de cada modelo, de forma granular, como um indicador de confiança. A premissa é que as previsões de um modelo devem ser mais valorizadas nas categorias de entidade em que ele demonstrou maior proficiência.

O processo requer, como pré-requisito, uma análise de desempenho detalhada de cada um dos $n = 7$ modelos base. Para cada modelo $M_i$ e para cada classe de entidade $c_k$, o *F1-Score* específico, $F(M_i, c_k)$, foi calculado e armazenado. O algoritmo de *ensemble* então opera da seguinte forma:

1. Para um dado texto, são coletadas todas as previsões de entidades únicas, definidas pela tupla (`índice_inicial, índice_final, rótulo`);

2. Para cada entidade única prevista, o sistema identifica o subconjunto de modelos que concordaram com essa previsão;

3. Em seguida, calcula-se a pontuação de confiança para aquela entidade. Essa pontuação corresponde à média aritmética simples dos F1-Scores (específicos daquela classe) dos modelos que a previram;

4. A entidade é mantida na saída final somente se sua pontuação de confiança for superior ao limiar $\tau$ predefinido.

Formalmente, seja $e$ uma entidade única prevista com o rótulo $c_e$. Seja $V(e)$ o conjunto de modelos que previram $e$. A pontuação de confiança $S_C(e)$ da entidade $e$ é calculada pela Equação 4.7.

$$S_C(e) = \frac{1}{|V(e)|} \sum_{M_i \in V(e)} F(M_i, c_e) \tag{4.7}$$

A decisão final é manter a entidade $e$ se $S_C(e) > \tau$. Para este trabalho, o limiar foi definido empiricamente como $\tau = 0.8$, buscando um equilíbrio que mantenha apenas as entidades com alta confiabilidade. Esta abordagem visa aprimorar a precisão do *ensemble*, valorizando as previsões apoiadas por modelos que são historicamente mais eficazes para aquela classe específica de entidade.

## 4.4 Critérios e Métricas de Avaliação

A avaliação de todos os modelos base e das estratégias de *ensemble* descritas neste capítulo foi realizada de forma sistemática e quantitativa. Conforme estabelecido na fundamentação teórica e em alinhamento com a ferramenta *spaCy* utilizada, o critério adotado foi o de avaliação exata (*exact match*).

O desempenho de cada abordagem foi medido com base nas três métricas padrão para a tarefa de REN: Precisão, Revocação e F1-Score. As definições formais e as equações de cada uma dessas métricas foram detalhadas previamente na Seção 2.4.1 deste documento. A aplicação dessas métricas ao conjunto de teste permitiu uma comparação justa e rigorosa entre os sete modelos individuais e as três estratégias de *ensemble*.

# Capítulo 5

# Resultados e Discussão

Este capítulo apresenta e discute os resultados obtidos a partir da aplicação da metodologia experimental detalhada no Capítulo 4. O foco da análise é avaliar o desempenho de sete modelos baseados na arquitetura *Transformer* e de três estratégias de *ensemble* na tarefa de Reconhecimento de Entidades Nomeadas (REN) no *corpus* de convênios do Diário Oficial da União.

A apresentação dos resultados segue uma narrativa lógica, partindo de uma visão quantitativa global, passando por uma análise granular detalhada e por uma análise qualitativa. A estrutura do capítulo está organizada da seguinte forma:

- Na Seção 5.1, discute-se o desempenho geral dos modelos e *ensembles*, utilizando métricas globais (*micro-average*) para oferecer um panorama da eficácia das abordagens propostas;
- Na Seção 5.2, a análise é aprofundada para o nível de cada categoria de entidade. Nesta etapa, investigam-se o *F1-Score*, a Precisão e a Revocação, além de examinar o comportamento dos modelos tanto em classes de alta disponibilidade de dados quanto naquelas que apresentaram desempenho crítico ou falhas totais;
- A Seção 5.3 apresenta uma análise qualitativa de erros e acertos. São apresentados casos reais onde as estratégias de *ensemble* demonstraram capacidade de corrigir omissões dos modelos individuais e, inclusive, superar a anotação original do gabarito, aumentando a cobertura na extração de entidades;
- A Seção 5.4 apresenta uma análise de sensibilidade, investigando como variações nos parâmetros e critérios de votação do *ensemble* impactam o resultado final;
- A Seção 5.5 encerra o capítulo com uma análise de custo computacional, comparando o esforço exigido para o ajuste fino (*fine-tuning*) com o ganho de desempenho efetivo proporcionado pela combinação dos modelos.

## 5.1 Desempenho Geral dos Modelos e Ensembles

Após a etapa de ajuste fino dos sete modelos base e da aplicação das três estratégias de combinação, foram conduzidos os experimentos para avaliar a eficácia de cada abordagem. Esta seção apresenta os resultados globais de desempenho, oferecendo um panorama da capacidade dos modelos e *ensembles* na tarefa de REN.

Em tarefas de classificação com múltiplas classes, como é o caso do REN neste trabalho, as métricas de desempenho globais podem ser calculadas por diferentes estratégias de agregação. Os resultados aqui reportados utilizam a abordagem de média *micro-average*, padrão adotado pela ferramenta de avaliação empregada [40]. Nesta abordagem, os totais de verdadeiros positivos, falsos positivos e falsos negativos de todas as classes de entidades são somados antes de se calcular a métrica final. Uma característica importante da *micro-average* é conferir um peso maior às classes mais populosas [53]. Dado o desbalanceamento de classes identificado no *corpus* (Seção 4.1.5), essa métrica geral tende a refletir o desempenho das entidades com o maior número de anotações no *corpus*. Essa particularidade reforça a necessidade de uma análise granular por entidade, realizada na Seção 5.2.

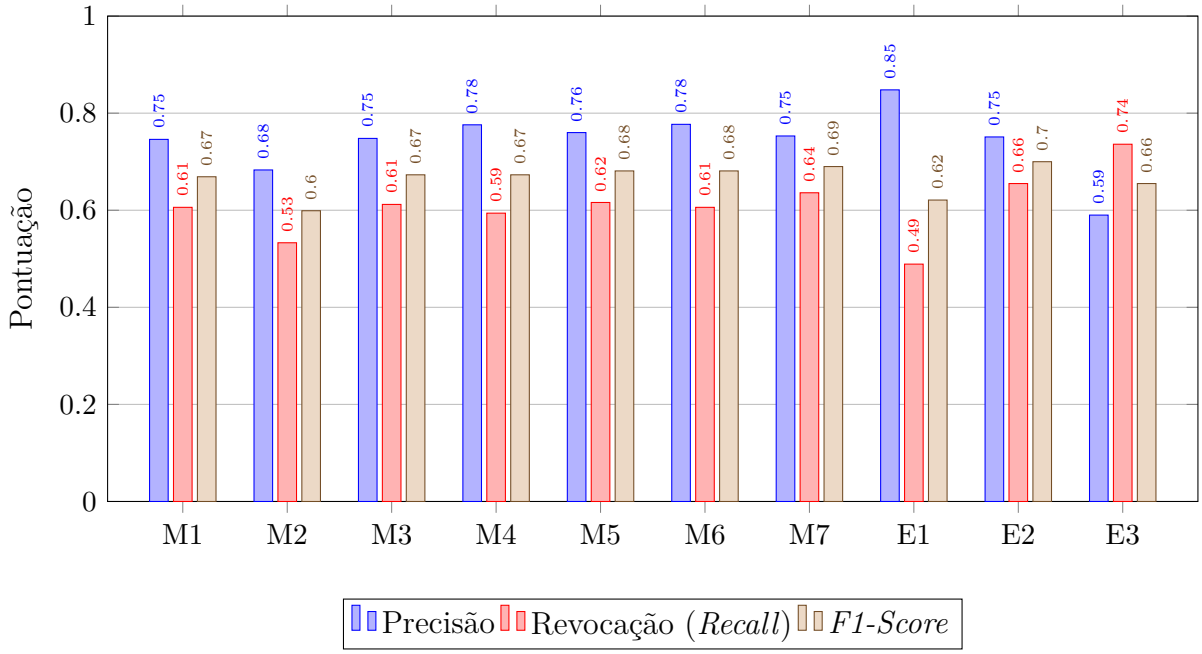
O desempenho consolidado de todos os modelos individuais e das estratégias de *ensemble* é apresentado na Tabela 5.1 e visualizado no gráfico da Figura 5.1. Para facilitar a análise, a identificação dos modelos (M1-M7) e dos *ensembles* (E1-E3) pode ser consultada, respectivamente, nas Tabelas 4.5 e 4.7.

Tabela 5.1: Resultados consolidados para os modelos individuais e as estratégias de *ensemble*. O melhor resultado em cada métrica está destacado em negrito.

| **Modelo / Estratégia** | **Precisão** | **Revocação (*Recall*)** | **F1-score** |
|---|---|---|---|
| M1 - [129] | 0,746 | 0,606 | 0,669 |
| M2 - [88] | 0,683 | 0,533 | 0,599 |
| M3 - [130] | 0,748 | 0,612 | 0,673 |
| M4 - [91] | 0,776 | 0,594 | 0,673 |
| M5 - [131] | 0,760 | 0,616 | 0,681 |
| M6 - [121] | 0,777 | 0,606 | 0,681 |
| M7 - [132] | 0,753 | 0,636 | 0,690 |
| E1 - Votação do Melhor Modelo | **0,848** | 0,489 | 0,621 |
| E2 - Votação por Maioria | 0,751 | 0,655 | **0,700** |
| E3 - Votação Ponderada | 0,590 | **0,736** | 0,655 |

A análise dos modelos base revela uma variação considerável de desempenho, com a *F1-Score* oscilando entre 0,599 (M2) e 0,690 (M7). O modelo M7 (xlm-roberta-base) destacou-se como o modelo individual com melhor desempenho, estabelecendo a principal linha de base para a avaliação comparativa das estratégias de combinação.

Figura 5.1: Desempenho geral dos modelos individuais (M1-M7) e das estratégias de *ensemble* (E1-E3).




Ao examinar os *ensembles*, observa-se um evidente compromisso entre Precisão e Revocação. A estratégia E1 especializou-se em alta Precisão, alcançando 0,848, um valor 9,1% superior ao do modelo individual mais preciso (M6). Contudo, esse ganho resultou na menor Revocação do estudo (0,489), indicando que o E1 foi um classificador conservador, que acertou com alta confiança o que se propôs a classificar, mas falhou em identificar uma grande parcela das entidades existentes. Em oposição, a estratégia E3 maximizou a Revocação, atingindo 0,736, um aumento de 15,7% em relação ao melhor modelo individual (M7). Este resultado, porém, veio ao custo da menor Precisão do experimento (0,590), caracterizando o E3 como um classificador mais "liberal", que identificou muitas entidades corretamente, mas também cometeu um número elevado de erros de classificação (falsos positivos).

A estratégia E2 (Votação por Maioria) apresentou o desempenho mais equilibrado e, consequentemente, o mais robusto, alcançando a maior *F1-Score* do estudo, com valor de 0,700. Este método conseguiu ampliar a capacidade de identificação de entidades (Revocação) em relação aos modelos individuais, sem incorrer na perda de Precisão observada no E3, o que resultou no melhor balanço entre as duas métricas. Para quantificar essa melhoria e comparar as análises deste trabalho, utiliza-se o Percentual de Melhoria ($PM$), definido na Equação 5.1.

$$PM = \left( \frac{V_e - V_i}{V_i} \right) \times 100 \tag{5.1}$$

onde $V_e$ representa o valor da métrica para o *ensemble* e $V_i$ representa o valor da métrica para o modelo individual de referência.

Ao comparar a *F1-Score* do melhor *ensemble* (E2) com a do melhor modelo individual (M7), o ganho foi de 1,45%. Embora modesta em termos percentuais, essa melhoria pode ser significativa ao se observar os ganhos por categorias. É importante ressaltar que, por se tratar de uma média global (*micro-average*), este ganho agregado pode ocultar melhorias mais substanciais em categorias de entidades específicas, especialmente naquelas que são mais desafiadoras para os modelos individuais. Esta hipótese é investigada em detalhe na Seção 5.2. Ainda assim, o resultado global evidencia que a combinação de modelos, mesmo sem uma otimização exaustiva de seus hiperparâmetros na etapa de ajuste fino, é capaz de gerar um sistema de REN mais eficaz e confiável do que qualquer um de seus componentes isoladamente.

## 5.2 Análise de Desempenho por Categoria de Entidade

Após a avaliação do desempenho agregado, esta seção aprofunda a análise no nível de cada categoria de entidade indicada. A investigação granular é fundamental para validar a hipótese levantada na seção anterior: de que a modesta melhoria geral dos *ensembles* poderia ocultar ganhos mais expressivos em entidades específicas, mais desafiadoras para os modelos individuais.

Para fornecer uma visão panorâmica inicial de quais modelos e *ensembles* se destacaram com maior frequência, a Tabela 5.2 apresenta um resumo do desempenho, mostrando tanto a contagem absoluta quanto o percentual de vezes em que cada abordagem alcançou o melhor resultado em cada uma das três métricas de avaliação para as 26 categorias de entidades nomeadas.

A análise da Tabela 5.2 confirma, de forma clara, a especialização de cada estratégia de *ensemble*. Na métrica de Precisão, observa-se uma dominância absoluta da estratégia E1 (65,4% das categorias). Em contrapartida, na Revocação, a hegemonia pertence ao E3 (53,8% das categorias). Na avaliação da F1-Score, métrica que pondera o equilíbrio crítico entre Precisão e Revocação, a superioridade do *ensemble* E2 é menos evidente quando comparado a E1 na precisão e a E3 na revocação. Contudo, no *F1-Score*, o E2 consolidou-se como a melhor abordagem (34,6% das categorias). Este resultado é particularmente expressivo quando contrastado com o M4, o modelo individual mais competitivo, que liderou em apenas 11,5% dos casos. Em termos práticos, a estratégia de combinação E2

[1] A soma pode não totalizar 26 (ou 100%) caso haja empates entre duas ou mais abordagens para o melhor desempenho em uma mesma categoria de entidade.

Tabela 5.2: Contagem e percentual de categorias de entidades nomeadas em que cada abordagem apresentou o melhor desempenho. As legendas para M1-M7 e E1-E3 são detalhadas, respectivamente, nas Tabelas 4.5 e 4.7.

| **Abordagem** | **Precisão** | | **Revocação (*Recall*)** | | **F1-Score** | |
|---|---|---|---|---|---|---|
| | **Nº Cat.** | **(%)** | **Nº Cat.** | **(%)** | **Nº Cat.** | **(%)** |
| E1 | **17** | 65,4% | 0 | 0,0% | 0 | 0,0% |
| E2 | 0 | 0,0% | 7 | 26,9% | **9** | 34,6% |
| E3 | 0 | 0,0% | **14** | 53,8% | 3 | 11,5% |
| M1 | 1 | 3,8% | 0 | 0,0% | 2 | 7,7% |
| M2 | 0 | 0,0% | 0 | 0,0% | 0 | 0,0% |
| M3 | 1 | 3,8% | 1 | 3,8% | 1 | 3,8% |
| M4 | 2 | 7,7% | 0 | 0,0% | 3 | 11,5% |
| M5 | 1 | 3,8% | 1 | 3,8% | 2 | 7,7% |
| M6 | 1 | 3,8% | 2 | 7,7% | 2 | 7,7% |
| M7 | 0 | 0,0% | 1 | 3,8% | 1 | 3,8% |
| **Total de Vitórias**[1] | 23 | 88,5% | 26 | 100,0% | 23 | 88,5% |

triplicou a frequência de vitórias em relação ao melhor modelo isolado, demonstrando uma consistência e robustez muito superiores na generalização do aprendizado. Para compreender como essas especializações se manifestam, a análise a seguir utiliza os dados detalhados nas Tabelas 5.3, 5.4 e 5.5.

### Categorização da Quantidade de Anotações

Para fundamentar as análises de desempenho a seguir, é necessário estabelecer categorias para o volume absoluto de anotações disponível no *corpus* consolidado. A distribuição da quantidade de anotações de cada categoria de entidades nomeadas, apresentada na Tabela 4.4, evidencia um desbalanceamento acentuado. Para mitigar possíveis ambiguidades na correlação estabelecida nas análises realizadas entre a quantidade de anotações e os resultados obtidos, foram definidos três limiares de disponibilidade dos dados no *corpus*, baseados na ordem de grandeza das amostras:

- **Alta Disponibilidade ($N > 10^5$):** Entidades com mais de 100.000 anotações. Este grupo, composto por 12 classes (ex.: 'Tipo Ente Convenente' e 'Data Início Vigência');
- **Média Disponibilidade ($10^4 < N \leq 10^5$):** Entidades entre 10.000 e 100.000 anotações. Este grupo intermediário abrange 7 classes (ex.: 'Número Convênio' e 'Nome Município'), nas quais o aprendizado é viável, mas a variância entre modelos tende a ser maior;

- **Baixa Disponibilidade ($N \leq 10^4$):** Entidades com menos de 10.000 anotações. Este grupo crítico, formado por 8 classes (ex.: 'Situação Convênio' e 'Tipo Convenente'), representa um cenário de escassez de dados, o que impõe desafios significativos à generalização dos modelos.

### 5.2.1 Análise de Desempenho por *F1-Score*

A *F1-Score* é amplamente reconhecida na literatura como a métrica de referência para a comparação de sistemas de Extração de Informação, consolidada desde as conferências MUC e CoNLL [135]. Sua relevância reside no fato de ser a média harmônica entre a Precisão e a Revocação (métricas detalhadas individualmente na Seção 2.4), condensando em um único índice o equilíbrio entre a qualidade e a abrangência das extrações.

Neste contexto, a análise com base nesta métrica oferece uma visão ampla do desempenho dos modelos. Os resultados detalhados, apresentados na Tabela 5.3, revelam três padrões distintos de comportamento, que se correlacionam fortemente com o nível de disponibilidade dos dados (conforme a categorização definida na Seção 5.2) e com a complexidade das entidades no *corpus* de treinamento.

O primeiro padrão ocorre predominantemente em entidades de Alta Disponibilidade ($N > 10^5$) que possuem estrutura textual regular, como 'TIPO ENTE CONVENENTE' (2,2M de anotações) e 'DATA INÍCIO VIGÊNCIA' (2,1M de anotações). Nestes casos, os modelos individuais já atingem um desempenho robusto, com o M3 e o M4 obtendo, respectivamente, as melhores pontuações. A saturação de exemplos de treinamento permite que os modelos aprendam a tarefa de forma eficaz, reduzindo a margem para que a combinação de modelos ofereça melhorias significativas.

O verdadeiro impacto das técnicas de *ensemble*, no entanto, se manifesta no segundo padrão: entidades de Média Disponibilidade ($10^4 < N \leq 10^5$) ou classes de Alta Disponibilidade que apresentam maior complexidade semântica. Neste cenário, há dados suficientes para o aprendizado, mas a variância estrutural ainda impõe desafios aos modelos isolados. Conforme já indicado na Tabela 5.2, os *ensembles* superaram os melhores modelos individuais em 12 das 26 categorias. Um exemplo notável é a entidade 'NOME ÓRGÃO CONCEDENTE' (Média Disponibilidade), na qual o *ensemble* E2 alcançou uma *F1-Score* de 0,901, um ganho de 2,5% em relação ao melhor modelo individual (M3, com 0,879). Ganhos ainda mais expressivos ocorreram em classes próximas ao limiar inferior, como 'NÚMERO ORIGINAL' (25 mil anotações), na qual o E2 obteve uma melhoria de 10,75% sobre o M3. Outros casos relevantes em que os *ensembles* se mostraram superiores incluem 'NOME CONVENENTE', 'NÚMERO CONVÊNIO', 'UF', 'VALOR CONTRAPARTIDA' e 'NOME MUNICÍPIO'. Esses resultados validam a hipótese de que a

combinação de modelos é particularmente eficaz para aumentar a robustez e corrigir erros esporádicos em cenários mais desafiadores.

O terceiro padrão revela uma limitação clara do treinamento supervisionado em contextos de Baixa Disponibilidade ($N \leq 10^4$). Para entidades com volume crítico de anotações, como 'TIPO CONVENENTE' (191 anotações) e 'VALOR LIBERADO' (162 anotações), nenhum modelo, individual ou combinado, foi capaz de aprender a tarefa, resultando em *F1-Score* nulo. Isso reforça que, apesar da sofisticação das arquiteturas, um número mínimo de exemplos de treinamento permanece como pré-requisito indispensável para o aprendizado em tarefas de REN.

Tabela 5.3: F1-Score por categoria de entidade. O melhor resultado entre os modelos individuais (M1-M7) está sempre em **negrito**. Adicionalmente, se um *ensemble* supera essa pontuação, seu resultado é destacado em **negrito e sublinhado**.

| **Entidade** | **Qt. Anot.** | **M1** | **M2** | **M3** | **M4** | **M5** | **M6** | **M7** | **E1** | **E2** | **E3** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CÓDIGO ÓRGÃO SUPERIOR | 1.114 | 0,154 | 0,000 | 0,047 | **0,662** | 0,092 | 0,234 | 0,156 | 0,389 | 0,114 | 0,000 |
| OBJETO DO CONVÊNIO | 114.439 | **0,444** | 0,106 | 0,330 | 0,303 | 0,294 | 0,300 | 0,315 | 0,435 | 0,289 | 0,000 |
| VALOR CONVÊNIO | 171.283 | 0,388 | 0,268 | 0,328 | 0,363 | 0,317 | **0,475** | 0,261 | 0,452 | 0,313 | 0,088 |
| CÓDIGO ÓRGÃO CONCEDENTE | 17.774 | 0,528 | 0,517 | 0,563 | 0,548 | 0,550 | **0,775** | 0,525 | 0,566 | 0,572 | 0,510 |
| DATA ÚLTIMA LIBERAÇÃO | 8.763 | **0,282** | 0,000 | 0,165 | 0,220 | 0,216 | 0,206 | 0,229 | 0,198 | 0,219 | 0,000 |
| NÚMERO PROCESSO DO CONVÊNIO | 38.371 | 0,625 | 0,561 | 0,704 | 0,587 | **0,765** | 0,743 | 0,595 | 0,566 | 0,678 | 0,579 |
| NOME CONVENENTE | 149.594 | 0,676 | 0,452 | 0,597 | 0,651 | 0,606 | 0,590 | **0,693** | 0,536 | 0,622 | **0,700** |
| DATA FINAL VIGÊNCIA | 508.623 | 0,615 | 0,545 | 0,647 | 0,654 | 0,618 | 0,545 | **0,670** | 0,614 | 0,649 | 0,000 |
| VALOR ÚLTIMA LIBERAÇÃO | 1.275.212 | 0,352 | 0,077 | 0,369 | 0,337 | **0,406** | 0,401 | 0,275 | 0,104 | 0,395 | 0,000 |
| DATA INÍCIO VIGÊNCIA | 2.190.278 | 0,762 | 0,735 | 0,753 | **0,782** | 0,778 | 0,741 | 0,777 | 0,747 | 0,772 | 0,766 |
| NOME ÓRGÃO SUPERIOR | 1.495.540 | **0,945** | 0,916 | 0,939 | 0,923 | 0,938 | 0,939 | 0,933 | 0,915 | 0,940 | **0,948** |
| NÚMERO CONVÊNIO | 92.891 | 0,709 | 0,696 | 0,708 | 0,712 | 0,711 | 0,714 | **0,719** | 0,686 | 0,717 | **0,722** |
| TIPO ENTE CONVENENTE | 2.274.833 | 0,951 | 0,918 | **0,952** | 0,946 | 0,944 | 0,942 | 0,950 | 0,947 | 0,951 | 0,947 |

Continua na próxima página

**Tabela 5.3 – continuação**

| Entidade | Qt. Anot. | M1 | M2 | M3 | M4 | M5 | M6 | M7 | E1 | E2 | E3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| TIPO INSTRUMENTO | 3.059 | 0,782 | 0,000 | 0,843 | 0,768 | 0,865 | **0,867** | 0,793 | 0,789 | **0,869** | 0,794 |
| CÓDIGO CONVENENTE | 102.837 | 0,665 | 0,677 | 0,692 | 0,731 | 0,751 | 0,751 | **0,753** | 0,625 | **0,756** | 0,754 |
| UF | 1.361.105 | 0,743 | 0,725 | 0,752 | 0,750 | 0,749 | 0,758 | **0,760** | 0,726 | **0,763** | 0,720 |
| CÓDIGO UG CONCEDENTE | 301.681 | 0,523 | 0,316 | 0,646 | 0,691 | 0,681 | 0,686 | **0,689** | 0,666 | **0,696** | 0,413 |
| VALOR CONTRAPARTIDA | 137.369 | 0,512 | 0,498 | 0,581 | 0,583 | 0,608 | **0,636** | 0,601 | 0,440 | **0,649** | 0,538 |
| NOME ÓRGÃO CONCEDENTE | 66.591 | 0,860 | 0,704 | **0,879** | 0,745 | 0,819 | 0,876 | 0,876 | 0,675 | **0,901** | 0,900 |
| NOME MUNICÍPIO | 68.081 | 0,531 | 0,323 | 0,541 | 0,523 | **0,563** | 0,558 | 0,560 | 0,515 | **0,578** | 0,000 |
| NOME UG CONCEDENTE | 1.383 | **0,581** | 0,000 | 0,531 | 0,371 | 0,441 | 0,197 | 0,211 | 0,418 | **0,598** | 0,421 |
| NÚMERO ORIGINAL | 25.771 | 0,359 | 0,325 | **0,400** | 0,381 | 0,369 | 0,393 | 0,386 | 0,301 | **0,443** | 0,000 |
| VALOR LIBERADO | 162 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| TIPO CONVENENTE | 191 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| SITUAÇÃO CONVÊNIO | 4.004 | 0,000 | 0,000 | 0,000 | **0,171** | 0,000 | 0,143 | 0,000 | 0,000 | 0,000 | 0,000 |
| CÓDIGO SIAFI MUNICÍPIO | 23.166 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |

### 5.2.2 Análise de Desempenho por Revocação (*Recall*)

A Revocação, também referida na literatura como cobertura, reflete o quão abrangente um sistema é em suas extrações [135]. No contexto desta pesquisa, mede-se a capacidade dos modelos e *ensembles* de identificar todas as entidades relevantes presentes nos convênios, penalizando as omissões (falsos negativos). Conforme a Tabela 5.2 já antecipou, é na Revocação que as estratégias de *ensemble* demonstram sua força de forma mais expressiva, superando os modelos individuais em 21 das 26 categorias de entidades (80,7% do total). O *ensemble* E3, em particular, destacou-se nesta métrica, obtendo o melhor desempenho global em 14 categorias.

A Tabela 5.4 detalha esses resultados. Observa-se que os *ensembles* E2 e E3 frequentemente alcançam valores de Revocação muito altos e atingem a pontuação máxima de 1,000 nas entidades 'CÓDIGO ÓRGÃO SUPERIOR' (para o E2) e 'NOME UG CONCEDENTE' (para o E3). Esse desempenho indica que, para essas categorias, as estratégias de combinação foram capazes de recuperar a totalidade dos casos positivos, reduzindo drasticamente a ocorrência de falsos negativos.

O maior ganho ocorreu na entidade 'NOME UG CONCEDENTE', que possui apenas 1.383 anotações. Nesta categoria, o *ensemble* E3 alcançou uma Revocação perfeita de 1,000, o que representa um ganho de 27,5% em relação ao melhor modelo individual (M1, com 0,784). Este resultado evidencia o potencial dos *ensembles* para maximizar a cobertura de detecção mesmo em entidades de Baixa Disponibilidade ($N \leq 10^4$). No entanto, vale ressaltar que um valor de Revocação tão elevado em uma classe com poucos dados pode indicar um comportamento excessivamente "liberal"do classificador. Esse fenômeno de *trade-off*, em que a maximização da Revocação pode ocorrer ao custo da Precisão, é explorado na seção seguinte. Em contraste, o menor ganho observado ocorreu para a entidade 'NÚMERO CONVÊNIO', na qual o E3 (0,571) superou o melhor modelo individual, M7 (0,570), por uma margem mínima, indicando que, para esta entidade, os modelos individuais já possuíam um desempenho consolidado.

Tabela 5.4: Revocação (*Recall*) por categoria de entidade. O melhor resultado entre os modelos individuais (M1-M7) está sempre em **negrito**. Adicionalmente, se um *ensemble* supera essa pontuação, seu resultado é destacado em **negrito e sublinhado**.

| **Entidade** | **Qt. Anot.** | **M1** | **M2** | **M3** | **M4** | **M5** | **M6** | **M7** | **E1** | **E2** | **E3** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CÓDIGO ÓRGÃO SUPERIOR | 1.114 | 0,875 | 0,000 | 0,667 | 0,797 | **1,000** | **1,000** | **1,000** | 0,840 | **1,000** | 0,000 |
| CÓDIGO ÓRGÃO CONCEDENTE | 17.774 | 0,768 | 0,927 | **0,977** | 0,727 | 0,830 | 0,751 | 0,715 | 0,554 | 0,956 | 0,974 |
| NÚMERO CONVÊNIO | 92.891 | 0,554 | 0,535 | 0,552 | 0,558 | 0,554 | 0,561 | **0,570** | 0,523 | 0,561 | **0,571** |
| CÓDIGO CONVENENTE | 102.837 | 0,509 | 0,539 | 0,541 | 0,593 | 0,615 | 0,618 | **0,635** | 0,456 | 0,620 | **0,636** |
| TIPO ENTE CONVENENTE | 2.274.833 | 0,943 | **0,946** | 0,942 | 0,939 | 0,937 | 0,937 | 0,943 | 0,923 | 0,943 | **0,949** |
| NOME ÓRGÃO SUPERIOR | 1.495.540 | **0,900** | 0,850 | 0,894 | 0,863 | 0,885 | 0,888 | 0,878 | 0,843 | 0,887 | **0,905** |
| NÚMERO PROCESSO DO CONVÊNIO | 38.371 | **0,775** | 0,740 | 0,719 | 0,578 | 0,647 | 0,639 | 0,522 | 0,623 | 0,738 | **0,782** |
| DATA INÍCIO VIGÊNCIA | 2.190.278 | **0,829** | 0,671 | 0,772 | 0,724 | 0,791 | 0,728 | 0,755 | 0,629 | 0,765 | **0,846** |
| CÓDIGO UG CONCEDENTE | 301.681 | 0,645 | 0,551 | **0,804** | 0,758 | 0,766 | 0,801 | 0,739 | 0,580 | 0,809 | **0,836** |
| UF | 1.361.105 | 0,650 | 0,599 | 0,643 | 0,624 | **0,669** | 0,639 | 0,651 | 0,574 | 0,644 | **0,720** |
| NOME CONVENENTE | 149.594 | 0,553 | 0,318 | 0,454 | 0,514 | 0,462 | 0,435 | **0,570** | 0,375 | 0,472 | **0,619** |
| TIPO INSTRUMENTO | 3.059 | 0,685 | 0,000 | 0,820 | 0,696 | 0,817 | **0,834** | 0,701 | 0,658 | 0,825 | **0,913** |
| NOME ÓRGÃO CONCEDENTE | 66.591 | 0,815 | 0,616 | 0,818 | 0,634 | 0,748 | **0,830** | 0,818 | 0,515 | 0,883 | **0,936** |

Continua na próxima página

**Tabela 5.4 – continuação**

| Entidade | Qt. Anot. | M1 | M2 | M3 | M4 | M5 | M6 | M7 | E1 | E2 | E3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| VALOR CONTRAPARTIDA | 137.369 | 0,434 | 0,350 | 0,514 | 0,482 | 0,635 | **0,651** | 0,522 | 0,288 | 0,618 | **0,753** |
| NOME UG CONCEDENTE | 1.383 | **0,784** | 0,000 | 0,591 | 0,370 | 0,540 | 0,123 | 0,135 | 0,342 | 0,833 | **1,000** |
| VALOR CONVÊNIO | 171.283 | 0,408 | 0,303 | 0,381 | 0,386 | 0,400 | 0,394 | **0,447** | 0,335 | 0,452 | **0,678** |
| NOME MUNICÍPIO | 68.081 | **0,524** | 0,298 | 0,490 | 0,438 | 0,478 | 0,465 | 0,486 | 0,380 | **0,525** | 0,000 |
| DATA FINAL VIGÊNCIA | 508.623 | 0,583 | 0,476 | **0,653** | 0,623 | 0,659 | 0,577 | 0,647 | 0,466 | **0,684** | 0,000 |
| NÚMERO ORIGINAL | 25.771 | 0,262 | **0,326** | 0,291 | 0,280 | 0,267 | 0,288 | 0,296 | 0,184 | **0,360** | 0,000 |
| DATA ÚLTIMA LIBERAÇÃO | 8.763 | **0,588** | 0,000 | 0,125 | 0,321 | 0,400 | 0,500 | 0,500 | 0,179 | **0,700** | 0,000 |
| VALOR ÚLTIMA LIBERAÇÃO | 1.275.212 | 0,285 | 0,279 | 0,385 | 0,271 | 0,370 | **0,436** | 0,274 | 0,114 | **0,547** | 0,000 |
| OBJETO DO CONVÊNIO | 114.439 | **0,480** | 0,268 | 0,441 | 0,344 | 0,321 | 0,425 | 0,429 | 0,418 | **0,631** | 0,000 |
| VALOR LIBERADO | 162 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| TIPO CONVENENTE | 191 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| SITUAÇÃO CONVÊNIO | 4.004 | 0,000 | 0,000 | 0,000 | 0,250 | 0,000 | **0,400** | 0,000 | 0,000 | 0,000 | 0,000 |
| CÓDIGO SIAFI MUNICÍPIO | 23.166 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |

### 5.2.3 Análise de Desempenho por Precisão

A Precisão reflete a qualidade das extrações realizadas pelo sistema, indicando quantas dessas extrações estão, de fato, corretas [135]. Uma alta precisão indica que o modelo é confiável ao classificar um termo como entidade, com baixa taxa de falsos positivos. Nesta métrica, a estratégia de *ensemble* E1, que seleciona as predições do modelo individual com melhor desempenho, demonstrou uma clara superioridade. Conforme a Tabela 5.2, o E1 obteve a melhor Precisão em 17 das 26 categorias (65,4%), superando as outras abordagens neste quesito.

A Tabela 5.5 detalha este desempenho. O E1 alcançou valores de Precisão expressivos, muitas vezes próximos da perfeição, em diversas categorias, como 'NÚMERO CONVÊNIO' (0,999), 'NOME ÓRGÃO SUPERIOR' (0,999) e 'NOME ÓRGÃO CONCEDENTE' (0,980). Este resultado sugere que a estratégia E1 atua como um classificador "conservador": ao isolar o modelo mais apto para aquela categoria específica, evita-se a introdução de ruído que poderia advir da votação de modelos menos competentes, minimizando, assim, a ocorrência de erros de inserção (falsos positivos).

O ganho mais notável do E1 ocorreu na entidade 'OBJETO DO CONVÊNIO'. O melhor modelo individual (M1) obteve uma Precisão de 0,413, enquanto o E1 elevou essa pontuação para 0,454, um ganho de 9,9%. Este é um resultado significativo, dado que esta entidade, apesar de possuir alta disponibilidade de dados, apresenta alta complexidade devido à sua natureza textual longa e não estruturada. Em contraste, o menor ganho percentual ocorreu em 'TIPO ENTE CONVENENTE', onde o E1 (0,971) superou, por uma margem estreita, o melhor modelo individual, M3 (0,962).

É fundamental, no entanto, contextualizar o sucesso do E1. A sua alta Precisão é, em grande parte, o reflexo de uma estratégia que sacrifica a abrangência. Como observado na seção anterior, o E1 frequentemente resulta em baixa Revocação. Ao confiar estritamente no "melhor especialista" e ignorar o consenso majoritário (que poderia resgatar entidades difíceis), o E1 perde muitas entidades (aumentando os falsos negativos) para garantir que as poucas classificadas estejam corretas. Esse comportamento ilustra o clássico *trade-off* entre Precisão e Revocação, reforçando a importância da *F1-Score* como a métrica de equilíbrio para a avaliação global do sistema.

Tabela 5.5: Precisão por categoria de entidade. O melhor resultado entre os modelos individuais (M1-M7) está sempre em **negrito**. Adicionalmente, se um *ensemble* supera essa pontuação, seu resultado é destacado em **negrito e sublinhado**.

| **Entidade** | **Qt. Anot.** | **M1** | **M2** | **M3** | **M4** | **M5** | **M6** | **M7** | **E1** | **E2** | **E3** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| VALOR ÚLTIMA LIBERAÇÃO | 1.275.212 | **0,460** | 0,045 | 0,354 | 0,447 | 0,451 | 0,371 | 0,276 | 0,095 | 0,309 | 0,000 |
| CÓDIGO ÓRGÃO SUPERIOR | 1.114 | 0,084 | 0,000 | 0,024 | **0,566** | 0,048 | 0,133 | 0,084 | 0,253 | 0,060 | 0,000 |
| NÚMERO PROCESSO DO CONVÊNIO | 38.371 | 0,524 | 0,452 | 0,690 | 0,596 | **0,935** | 0,886 | 0,692 | 0,518 | 0,628 | 0,460 |
| CÓDIGO ÓRGÃO CONCEDENTE | 17.774 | 0,402 | 0,358 | 0,396 | 0,439 | 0,411 | **0,801** | 0,414 | 0,579 | 0,408 | 0,346 |
| DATA ÚLTIMA LIBERAÇÃO | 8.763 | 0,185 | 0,000 | **0,241** | 0,167 | 0,148 | 0,130 | 0,148 | 0,222 | 0,130 | 0,000 |
| NOME ÓRGÃO SUPERIOR | 1.495.540 | 0,996 | 0,992 | 0,989 | 0,991 | **0,997** | 0,996 | 0,996 | **0,999** | 0,998 | 0,994 |
| NÚMERO CONVÊNIO | 92.891 | 0,983 | **0,996** | 0,986 | 0,982 | 0,991 | 0,981 | 0,973 | **0,999** | 0,994 | 0,981 |
| TIPO ENTE CONVENENTE | 2.274.833 | 0,959 | 0,892 | **0,962** | 0,953 | 0,951 | 0,948 | 0,957 | **0,971** | 0,958 | 0,944 |
| NOME CONVENENTE | 149.594 | 0,869 | 0,781 | 0,869 | 0,887 | 0,884 | **0,917** | 0,883 | **0,940** | 0,912 | 0,806 |
| CÓDIGO CONVENENTE | 102.837 | 0,958 | 0,911 | 0,961 | 0,954 | **0,964** | 0,958 | 0,926 | **0,995** | 0,971 | 0,926 |
| NOME ÓRGÃO CONCEDENTE | 66.591 | 0,910 | 0,821 | **0,949** | 0,902 | 0,905 | 0,927 | 0,943 | **0,980** | 0,918 | 0,866 |
| UF | 1.361.105 | 0,867 | 0,918 | 0,905 | **0,940** | 0,851 | 0,931 | 0,915 | **0,988** | 0,936 | 0,720 |
| TIPO INSTRUMENTO | 3.059 | 0,913 | 0,000 | 0,867 | 0,856 | **0,918** | 0,903 | 0,913 | **0,985** | 0,918 | 0,703 |

Continua na próxima página

**Tabela 5.5 – continuação**

| **Entidade** | **Qt. Anot.** | **M1** | **M2** | **M3** | **M4** | **M5** | **M6** | **M7** | **E1** | **E2** | **E3** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| DATA INÍCIO VIGÊNCIA | 2.190.278 | 0,705 | 0,812 | 0,734 | **0,850** | 0,765 | 0,754 | 0,799 | **0,919** | 0,778 | 0,700 |
| VALOR CONTRAPARTIDA | 137.369 | 0,623 | **0,859** | 0,667 | 0,739 | 0,584 | 0,622 | 0,709 | **0,929** | 0,683 | 0,419 |
| OBJETO DO CONVÊNIO | 114.439 | **0,413** | 0,066 | 0,264 | 0,271 | 0,272 | 0,232 | 0,249 | **0,454** | 0,187 | 0,000 |
| NOME UG CONCEDENTE | 1.383 | 0,462 | 0,000 | 0,483 | 0,373 | 0,373 | **0,487** | 0,479 | **0,538** | 0,466 | 0,267 |
| NOME MUNICÍPIO | 68.081 | 0,538 | 0,354 | 0,604 | 0,651 | 0,686 | **0,698** | 0,662 | **0,800** | 0,645 | 0,000 |
| VALOR CONVÊNIO | 171.283 | 0,371 | 0,240 | 0,288 | 0,342 | 0,262 | **0,598** | 0,184 | **0,694** | 0,239 | 0,047 |
| CÓDIGO UG CONCEDENTE | 301.681 | 0,440 | 0,221 | 0,539 | 0,634 | 0,613 | 0,599 | **0,645** | **0,781** | 0,610 | 0,274 |
| DATA FINAL VIGÊNCIA | 508.623 | 0,651 | 0,638 | 0,641 | 0,687 | 0,581 | 0,516 | **0,695** | **0,901** | 0,618 | 0,000 |
| NÚMERO ORIGINAL | 25.771 | 0,569 | 0,325 | **0,640** | 0,596 | 0,599 | 0,617 | 0,551 | **0,830** | 0,575 | 0,000 |
| VALOR LIBERADO | 162 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| TIPO CONVENENTE | 191 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| SITUAÇÃO CONVÊNIO | 4.004 | 0,000 | 0,000 | 0,000 | **0,130** | 0,000 | 0,087 | 0,000 | 0,000 | 0,000 | 0,000 |
| CÓDIGO SIAFI MUNICÍPIO | 23.166 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |

### 5.2.4 Desempenho em Entidades de Alta Disponibilidade

A análise da Tabela 5.3 corrobora a discussão sobre o efeito da *micro-average*. As entidades classificadas como de Alta Disponibilidade no *corpus* (conforme definido na Seção 5.2), como 'TIPO ENTE CONVENENTE', 'NOME ÓRGÃO SUPERIOR' e 'DATA INÍCIO VIGÊNCIA', são aquelas que, em geral, apresentam os maiores valores de *F1-Score*.

Para a entidade 'TIPO ENTE CONVENENTE', por exemplo, quase todos os modelos e *ensembles* atingiram *F1-Score* superiores a 0,940. Este alto desempenho é resultado direto do vasto suporte de treinamento ($N > 2,2$ milhões), que permitiu obter valores consistentemente elevados tanto em Precisão quanto em Revocação. Isso indica que os padrões linguísticos dessas entidades foram aprendidos com eficácia. Nestes casos de saturação de dados, os modelos individuais já atingem um nível de desempenho muito alto, deixando pouco espaço para melhorias significativas por meio da combinação de modelos.

### 5.2.5 O Impacto dos Ensembles em Entidades Desafiadoras

Nesta análise, "Entidades Desafiadoras" são aquelas que impõem barreiras significativas à generalização dos modelos, seja pela escassez de exemplos de treinamento (Baixa ou Média Disponibilidade) ou pela alta variabilidade estrutural e semântica do texto (campos livres e não padronizados). É justamente nestes cenários de maior complexidade que o valor das estratégias de combinação se manifesta, o que valida a hipótese central deste trabalho. A agregação das decisões de múltiplos modelos resultou em ganhos e na otimização do balanço crítico entre Precisão e Revocação.

Um caso exemplar no espectro de Média Disponibilidade é a entidade 'NOME ÓRGÃO CONCEDENTE'. O melhor modelo individual (M3) alcançou uma *F1-Score* de 0,879, com uma alta Precisão (0,949) mas uma Revocação inferior (0,818). O *ensemble* E2, por sua vez, elevou a *F1-Score* para 0,901. Essa melhoria foi impulsionada por um aumento substancial na Revocação para 0,883, mantendo a Precisão em patamares elevados (0,918). Ou seja, o E2 aprendeu a recuperar mais entidades corretas do que qualquer modelo isolado, atingindo o melhor equilíbrio entre as métricas.

Um padrão similar ocorreu em cenários de Baixa Disponibilidade, como na entidade 'NOME UG CONCEDENTE' (apenas 1.383 anotações). O melhor *F1-Score* individual foi de 0,581 (M1), enquanto o *ensemble* E2 atingiu 0,598. O ganho de 2,9% sobre o melhor modelo individual, embora pareça modesto em números absolutos, revela um comportamento importante: o E2 obteve um aumento expressivo de Revocação (0,833 vs. 0,784 do M1), demonstrando que a combinação foi mais eficaz em encontrar as poucas instâncias dispersas desta classe.

Contudo, os *ensembles* não são uma solução universal, especialmente quando a complexidade advém da estrutura textual e não apenas da falta de dados. Para a entidade 'OBJETO DO CONVÊNIO', que possui Alta Disponibilidade, mas é caracterizada por ser um texto de longa extensão e alta variabilidade, o melhor desempenho geral permaneceu com um modelo individual (M1, *F1-Score* de 0,444). A estratégia E1, especialista em Precisão, alcançou uma pontuação similar (0,435), melhorando a Precisão do M1 (0,454 vs. 0,413) e mantendo uma Revocação comparável. Já o E2, que promove a concordância majoritária, teve um desempenho insatisfatório (0,289), pois sua Revocação alta (0,631) foi acompanhada de Precisão muito baixa (0,187). Isso sugere que, para entidades muito heterogêneas, a discordância entre os modelos base é tão acentuada que a votação majoritária acaba por introduzir um excesso de ruído (falsos positivos), degradando o desempenho global.

### 5.2.6 Análise de Entidades de Baixo Desempenho e Falhas Totais

A análise por categoria expõe as severas limitações dos modelos em cenários de Baixa Disponibilidade ($N \leq 10^4$). Conforme observado na Tabela 5.3, entidades situadas nesta faixa, como 'SITUAÇÃO CONVÊNIO' e 'CÓDIGO ÓRGÃO SUPERIOR', apresentaram *F1-Score* muito baixos, evidenciando a dificuldade de convergência da aprendizagem quando o suporte de treinamento é reduzido.

Os casos mais críticos, no entanto, são as entidades que apresentaram desempenho consistentemente nulo em quase todos os modelos e *ensembles*: 'VALOR LIBERADO', 'TIPO CONVENENTE' e 'CÓDIGO SIAFI MUNICÍPIO'.

Uma análise detalhada do *corpus* (Tabela 4.4) revela que 'VALOR LIBERADO' (162 anotações) e 'TIPO CONVENENTE' (191 anotações) ocupam o limiar inferior da Baixa Disponibilidade. O desempenho nulo nestes casos é um exemplo clássico de *extreme data sparsity* (escassez extrema de dados), em que o número de exemplos é insuficiente para que arquiteturas complexas, como os *Transformers*, consigam extrair padrões generalizáveis. Já para a entidade 'CÓDIGO SIAFI MUNICÍPIO', a falha total sugere desafios adicionais, indicando, por exemplo, uma ambiguidade em relação a outros códigos numéricos. Mesmo possuindo um volume intermediário de dados, os modelos não conseguiram distinguir sua semântica.

Esses achados reforçam a premissa de que a qualidade e, sobretudo, a quantidade mínima de dados de treinamento continuam sendo fatores determinantes para o sucesso em tarefas de REN, impondo um teto de desempenho que nem mesmo a combinação de modelos consegue superar.

## 5.3 Análise Qualitativa de Acertos e Erros

Para além das métricas quantitativas, a análise qualitativa de exemplos específicos é fundamental para ilustrar o impacto prático das diferentes abordagens e validar a hipótese de que os modelos treinados são capazes de superar as limitações do *corpus* de treinamento (detalhada na Seção 4.1.5). Com base no Desempenho Geral dos Modelos e Ensembles (Seção 5.1), nesta Seção, são apresentados dois estudos de caso para comparar visualmente o gabarito original (anotado por correspondência exata), as predições do melhor modelo individual (M7) e as do *ensemble* de melhor desempenho geral (E2) na métrica *F1-Score*

### 5.3.1 Caso 1: Correção de Erro de Omissão e Superação do Gabarito

Uma análise aprofundada do exemplo apresentado nas Figuras 5.2a, 5.2b e 5.2c revela não apenas a superioridade do *ensemble*, mas também as limitações da própria anotação automática por correspondência exata usada para gerar o gabarito. Enquanto o gabarito original continha apenas 7 entidades para este texto, os modelos M7 e E2 foram capazes de inferir, a partir do contexto, a presença de entidades implicitamente corretas, como 'CÓDIGO UG CONCEDENTE' e 'DATA INÍCIO VIGÊNCIA' (inclusive associada à "Data de Assinatura"), demonstrando uma capacidade de generalização que supera a do método de anotação.

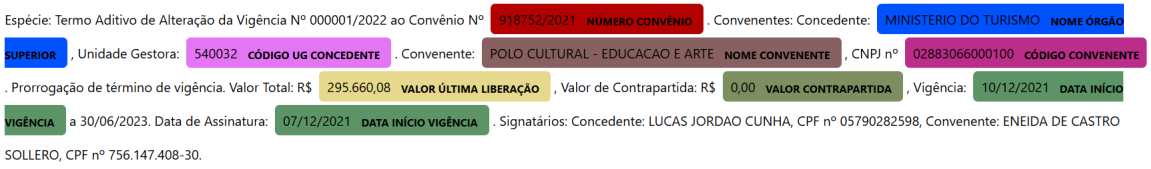
O modelo M7 (Figura 5.2b), embora tenha superado o gabarito (Figura 5.2a), cometeu um erro crucial de omissão ao não identificar a entidade 'DATA FINAL VIGÊNCIA' (com o valor '30/06/2023'). É neste ponto que o valor do *ensemble* E2 (Figura 5.2c) se manifesta de forma clara: ele não só manteve os acertos contextuais que o M7 fez, como também corrigiu a falha de omissão, identificando corretamente a 'DATA FINAL VIGÊNCIA'.

Portanto, este caso ilustra um cenário de dupla vitória para o *ensemble*: ele não apenas herda a capacidade dos modelos base de superar um gabarito imperfeito, mas também corrige erros críticos que até mesmo o melhor modelo individual comete. O *ensemble* atua como um sistema de verificação e complementação, resultando em uma extração de informações visivelmente mais completa e confiável, o que valida seu uso prático.

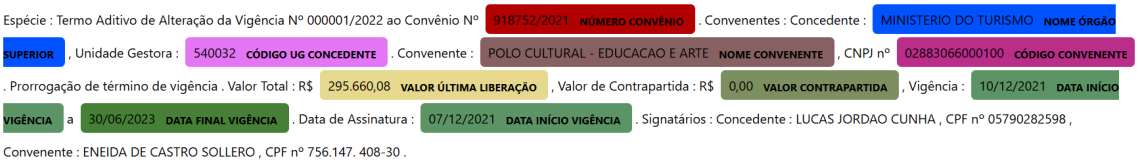
### 5.3.2 Caso 2: Aumento da Cobertura na Extração de Entidades

Para demonstrar a abrangência dos benefícios do *ensemble*, um segundo estudo de caso, centrado em um extrato de convênio, é analisado nas Figuras 5.3a, 5.3b e 5.3c. Este exemplo ilustra uma faceta distinta da contribuição oriunda da combinação de modelos: o aumento da cobertura na extração de informações.


(a) Gabarito original.




(b) Predições do modelo (M7), que supera o gabarito, mas omite a 'DATA FINAL VIGÊNCIA'.



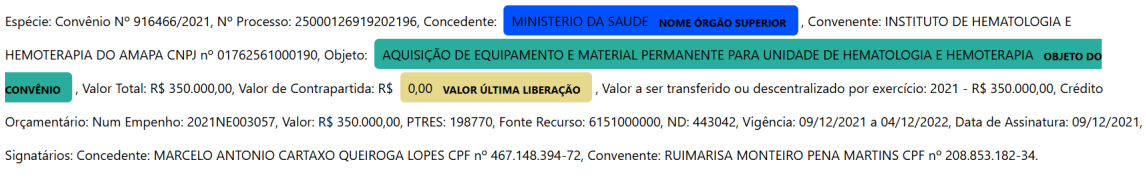

(c) Predições do *ensemble* E2, que corrige a omissão do M7.

Figura 5.2: Comparativo Caso 1.

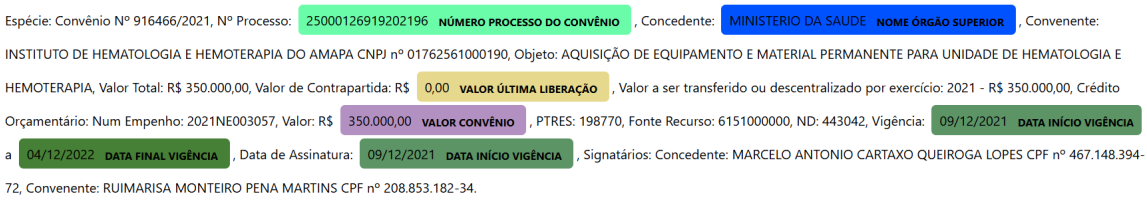
Similarmente ao caso anterior, tanto o M7 quanto o E2 demonstram uma capacidade de extração contextual superior ao gabarito gerado por correspondência exata, que identificou apenas três entidades neste texto. O modelo M7 (Figura 5.3b) foi bem-sucedido em identificar entidades estruturadas como datas e números de processo. No entanto, ele falhou em extrair informações fundamentais sobre as partes envolvidas, omitindo completamente o 'NÚMERO CONVÊNIO', o 'NOME CONVENENTE' e o respectivo 'CÓDIGO CONVENENTE'.

A contribuição do *ensemble* E2 (Figura 5.3c) é novamente clara e complementar à do primeiro caso. A abordagem de votação majoritária foi capaz de preencher as lacunas deixadas pelo M7, identificando corretamente todas as três entidades que o modelo individual havia perdido. É importante notar, contudo, que ambas as abordagens falharam em extrair certas entidades, como 'OBJETO DO CONVÊNIO', que é textualmente longa e complexa, o que indica uma limitação persistente para este tipo de entidade.

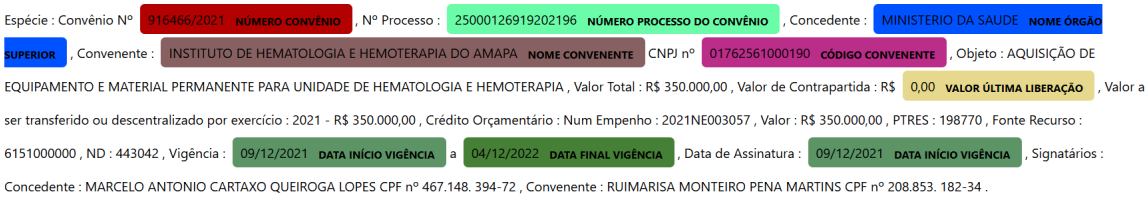
Entretanto, este exemplo demonstra a capacidade do *ensemble* de melhorar a Revocação em entidades-chave. Desta forma, o resultado final não é apenas mais correto, como no primeiro caso, mas significativamente mais completo, o que, em um cenário de aplicação real, pode ser de imenso valor para a análise de dados.




(a) Gabarito original, visivelmente incompleto.




(b) Predições do modelo (M7), que supera o gabarito, mas que omite as informações do convenente.




(c) Predições do *ensemble* E2, preenchendo as lacunas deixadas pelo M7.

Figura 5.3: Comparativo Caso 2.

## 5.4 Análise de Sensibilidade dos Parâmetros do Ensemble

Para validar a robustez das principais estratégias de *ensemble* (E1, E2 e E3) e compreender a escolha dos valores de seus parâmetros, foi realizada uma análise de sensibilidade. Esta análise consistiu na avaliação de configurações alternativas para cada uma das três técnicas de combinação, variando parâmetros-chave como o limiar de decisão, a métrica de ponderação e o critério de seleção do especialista. O objetivo é demonstrar empiricamente que as configurações escolhidas para E1, E2 e E3 são adequadas aos objetivos propostos neste trabalho, a fim de evidenciar que a combinação de modelos é uma estratégia capaz de aprimorar o desempenho e a robustez do REN no contexto de convênios.

A Tabela 5.6 consolida os resultados gerais (Precisão, Revocação e F1-Score) dos testes de sensibilidade, comparando-os às três estratégias principais, destacadas em negrito.

A análise dos resultados valida as escolhas metodológicas. Para a Votação por Maioria, os dados indicam que a configuração do E2 (limiar de 4 votos) é ótima. Um limiar mais permissivo (3 votos, Teste A) resultou em uma *F1-Score* significativamente menor (0,681), enquanto um limiar mais restritivo (6 votos, Teste B) também levou a uma leve queda na performance (0,696), o que justifica a escolha da maioria simples como o melhor balanço.

Para a Votação Ponderada, a análise revela um claro *trade-off* entre a Precisão e a

Tabela 5.6: Resultados da análise de sensibilidade dos parâmetros das estratégias de *ensemble*. As configurações escolhidas para o trabalho (E1, E2, E3) estão em negrito.

| ID do Teste | Estratégia | Parâmetro Chave | Precisão | Revocação | F1-Score |
|---|---|---|---|---|---|
| **Análise da Estratégia: Votação por Maioria** | | | | | |
| Teste A | Votação por Maioria | Limiar de 3/7 votos | 0,833 | 0,576 | 0,681 |
| **E2 (Escolhido)** | **Votação por Maioria** | **Limiar de 4/7 votos (maioria)** | **0,751** | **0,655** | **0,700** |
| Teste B | Votação por Maioria | Limiar de 6/7 votos | 0,702 | 0,690 | 0,696 |
| **Análise da Estratégia: Votação Ponderada** | | | | | |
| Teste C | Votação Ponderada (F1) | Limiar de 0.3 | 0,786 | 0,622 | 0,694 |
| Teste D | Votação Ponderada (F1) | Limiar de 0.5 | 0,720 | 0,667 | 0,693 |
| **E3 (Escolhido)** | **Votação Ponderada (F1)** | **Limiar de 0.8** | **0,590** | **0,736** | **0,655** |
| Teste E | Votação Ponderada (Recall) | Limiar de 0.5 | 0,720 | 0,667 | 0,693 |
| **Análise da Estratégia: Melhor Modelo** | | | | | |
| **E1 (Escolhido)** | **Max F1 por Entidade** | **Critério: F1-Score** | **0,848** | **0,489** | **0,621** |
| Teste F | Max Recall por Entidade | Critério: Revocação | 0,847 | 0,488 | 0,619 |

Revocação, controlado pelo limiar. Embora limiares mais baixos (0,3 e 0,5, Testes C e D) produzissem *F1-Score* geral mais alto (0,694 e 0,693), a configuração escolhida do E3 (limiar 0,8) cumpriu seu objetivo específico de maximizar a Revocação, alcançando o valor de 0,736, o mais alto entre todos os testes. Isso confirma que a escolha do E3 foi deliberada para criar um classificador especialista de alta cobertura. Essa característica torna a configuração do E3 particularmente útil para casos de uso como tarefas de auditoria, onde o objetivo principal é minimizar a perda de informações (falsos negativos) e garantir a identificação do maior conjunto possível de entidades candidatas para uma posterior revisão, mesmo que isso implique em uma maior taxa de falsos positivos. Além disso, o Teste E, que usou a Revocação como peso, não mostrou vantagem em relação ao uso da F1-Score.

Finalmente, ao comparar a estratégia do Melhor Modelo baseada em *F1-Score* (E1) com a seleção por Revocação (Teste F), observa-se equivalência prática nos resultados finais (0,621 vs 0,619). Embora a variação numérica seja marginal, a manutenção do critério de seleção baseado no *F1-Score* se justifica conceitualmente: por se tratar de uma média harmônica, essa métrica oferece um balanço mais seguro entre Precisão e Revocação, evitando a seleção de modelos que maximizem a recuperação de entidades às custas de um aumento excessivo de falsos positivos. Em suma, a análise de sensibilidade demonstra a estabilidade do método, indicando que a estratégia adotada no *ensemble* E1 é robusta e consistente com os objetivos globais do sistema.

Para fins de transparência e reprodutibilidade, a compilação completa dos experimentos realizados nesta análise de sensibilidade (incluindo as métricas de Precisão, Revocação e *F1-Score* discriminadas por categoria de entidade para cada configuração de limiar testada) encontra-se detalhada no Apêndice **A**. Esta documentação suplementar permite a verificação detalhada das decisões que fundamentaram a seleção dos parâmetros finais dos *ensembles* E1, E2 e E3.

## 5.5 Análise de Custo Computacional: Ajuste Fino versus Ensemble

Para validar a viabilidade prática da arquitetura proposta, foi realizada uma análise comparativa de custos computacionais entre a construção do *ensemble* e a busca exaustiva por hiperparâmetros (*fine-tuning* intensivo) em um único modelo isolado.

Como linha de base (*baseline*) para a medição de tempo, utilizou-se o modelo *Transformer* mais leve avaliado neste trabalho, o ALBERT Base v2 (*A Lite BERT*). Esta arquitetura destaca-se pelo baixo uso de memória, com apenas 11 milhões de parâmetros, devido à sua estratégia de compartilhamento de camadas [129]. O treinamento foi realizado na mesma infraestrutura de GPU utilizada nos experimentos principais.

Os registros de execução (*logs*) indicaram que um único ciclo de treinamento completo (13 épocas, $\approx$ 7.400 passos) deste modelo leve consumiu 2 horas e 17 minutos. A Tabela 5.7 projeta o custo estimado para tentar superar a performance do *ensemble* através de métodos tradicionais de otimização (*Grid Search*). A projeção contrasta as versões *Base* (utilizadas nesta pesquisa) com as *Large*, comuns no estado da arte, evidenciando a escalabilidade do custo.

Tabela 5.7: Comparativo de Custo Computacional Estimado: Otimização vs. Ensemble

| **Estratégia** | **Versão do Modelo** | **Qtd. Treinos** | **Tempo Estimado** | **Custo Relativo** |
|---|---|---|---|---|
| Treino Único (Baseline) | ALBERT Base (11M) | 1 | $\approx$ 2,3 horas | 1x (Base) |
| Treino Único (Pesado) | BERT Large (340M) | 1 | $\approx$ 7 a 10 horas* | 4x |
| Otimização (Grid Search) | ALBERT Base (11M) | 20 (combinações) | $\approx$ 46 horas | 20x |
| Otimização (Grid Search) | BERT Large (340M) | 20 (combinações) | $\approx$ 140 a 200 horas | >60x |
| **Ajuste do Ensemble** | **Qualquer Versão** | **0 (Pós-proc.)** | **< 1 minuto** | **Desprezível** |

Estimativa aproximada baseada na proporção de parâmetros e operações de ponto flutuante (FLOPs).

A análise demonstra que, embora este trabalho tenha adotado as versões *Base* dos modelos (ex.: BERT-base, RoBERTa-base), a tentativa de superar as métricas do *ensemble* por meio do refinamento individual já impõe um custo de dezenas de horas. Caso fossem utilizadas as versões *Large*, esse custo se tornaria proibitivo (podendo ultrapassar uma semana de processamento) para ganhos incrementais incertos.

Em contrapartida, a alteração de parâmetros do *ensemble* (como o ajuste de pesos de votação ou de limiares de confiança) ocorre estritamente na etapa de inferência e independe do tamanho do modelo neural subjacente. Portanto, modificar a estratégia de combinação é computacionalmente instantâneo, enquanto reconfigurar modelos robustos exige dias de processamento, o que confere à abordagem de *ensemble* uma relação custo-benefício superior.

### 5.5.1 Considerações sobre o Tempo de Inferência

É importante ressaltar que as métricas de tempo discutidas neste contexto devem ser interpretadas como estimativas aproximadas. A realização de uma medição rigorosa e isolada do tempo de execução não foi viável devido à natureza do ambiente experimental, que consiste em um servidor de acesso compartilhado (as características do servidor encontram-se na Seção 4.2.3). A concorrência variável por recursos de hardware durante os experimentos, somada a incompatibilidades técnicas das ferramentas de monitoramento com a arquitetura específica, introduz flutuações nas medições. Portanto, os valores indicados representam ordens de grandeza operacionais e não tempos absolutos de um ambiente dedicado.

Ressalva-se, ainda, que, diferentemente da etapa de treinamento, a fase de inferência no *ensemble* apresenta um custo computacional proporcional ao número de modelos utilizados ($N = 7$). Contudo, a análise de requisitos do domínio de processamento de atos oficiais do Diário Oficial da União indica que a operação ocorre majoritariamente em lote (*batch processing*). Desta forma, o sistema não exige respostas em tempo real (baixa latência), o que torna aceitável o custo computacional adicional.

Assim, neste cenário de aplicação, a penalidade temporal na inferência é tecnicamente admissível em favor da maximização da confiabilidade dos dados extraídos. Dada a natureza crítica dos convênios, prioriza-se a robustez da classificação em relação à velocidade de processamento individual dos documentos.

# Capítulo 6

# Considerações Finais

A transparência na administração pública e a fiscalização eficiente dos gastos governamentais dependem, cada vez mais, da capacidade de processar volumes massivos de dados não estruturados. Este trabalho partiu da premissa de que a aplicação de técnicas avançadas de Processamento de Linguagem Natural (PLN), especificamente o Reconhecimento de Entidades Nomeadas (REN), é fundamental para converter a opacidade textual dos Diários Oficiais em dados estruturados e acionáveis.

O objetivo central desta dissertação foi investigar e avaliar o impacto de técnicas de *ensemble*, aplicadas a um comitê de modelos baseados na arquitetura *Transformer*, como estratégia para aprimorar o desempenho e a robustez do Reconhecimento de Entidades Nomeadas (REN) em extratos de convênios públicos do DOU. Para viabilizar tal investigação, cumpriu-se inicialmente o objetivo de preencher uma lacuna na literatura por meio da construção de um *corpus* em larga escala via estratégia de anotação automática, conforme detalhado no Capítulo 4. A partir desta base de dados, foi possível avaliar o ajuste fino de sete modelos distintos e comparar diferentes estratégias de votação, permitindo uma análise aprofundada dos resultados frente aos desafios de desbalanceamento de classes do domínio.

Ao revisitar os objetivos específicos traçados na Seção 1.1, conclui-se que todos foram atingidos satisfatoriamente. A construção do *corpus*, o ajuste fino de sete modelos e a avaliação sistemática das estratégias de combinação permitiram não apenas responder às questões de pesquisa, mas também fornecer artefatos práticos à comunidade científica e ao projeto *Deep Vacuity* [12].

Para consolidar o encerramento deste estudo, o presente capítulo está organizado da seguinte forma: a Seção 6.1 apresenta uma síntese dos principais resultados empíricos obtidos; a Seção 6.2 destaca as contribuições científicas e práticas da pesquisa; a Seção 6.3 discute as limitações metodológicas e técnicas identificadas; e, por fim, a Seção 6.4 sugere caminhos para a continuidade e expansão desta investigação.

## 6.1 Síntese dos Resultados Obtidos

A análise detalhada dos experimentos (apresentada no Capítulo 5) permite extrair conclusões sólidas sobre a aplicação de *Transformers* e *ensembles* no domínio jurídico - administrativo brasileiro. O primeiro achado central é que, embora o melhor modelo individual (M7 - RoBERTa-Base) tenha estabelecido uma linha de base robusta, a estratégia de *ensemble* por Votação por Maioria (E2) demonstrou superioridade.

Conforme evidenciado na Tabela de Resultados Gerais (ver Tabela 5.1), o E2 apresentou o melhor desempenho equilibrado (F1-Score), superando os modelos individuais. A melhoria trazida pelo *ensemble*, embora numericamente modesta na métrica global, revelou-se qualitativamente expressiva na análise granular discriminada por entidade (ver Seção 5.2). O método E2 foi capaz de corrigir erros críticos de omissão e preencher lacunas de extração deixadas até mesmo pelo melhor modelo individual, aumentando, assim, a confiabilidade do sistema final.

Adicionalmente, a exploração das diferentes estratégias de combinação (E1, E2, E3) discutidas e apresentadas na Seção 4.3 tornou tangível o controle do *trade-off* entre Precisão e Revocação. Os dados demonstram ser possível configurar o sistema para objetivos distintos:

- **Cenário Conservador (E1):** Ideal para a automação de sistemas críticos, onde a tolerância a falsos positivos é baixa (alta precisão);
- **Cenário de Auditoria/Varredura (E3):** Focado em maximizar a cobertura (alta revocação), garantindo que nenhum indício de irregularidade seja ignorado, mesmo que isso implique maior esforço humano na revisão posterior.

Contudo, observou-se que o desempenho permanece fortemente correlacionado à frequência das classes no *corpus*. Entidades raras sofreram com a escassez de dados (*data sparsity*), indicando que nem mesmo a robustez dos *ensembles* é capaz de compensar plenamente a falta de exemplos de treinamento para categorias muito infrequentes.

## 6.2 Contribuições desta Pesquisa

As contribuições deste trabalho estendem-se além dos resultados experimentais, impactando tanto a metodologia de pesquisa em REN quanto a aplicação prática na gestão pública:

1. **Criação de Corpus de Larga Escala:** Uma importante contribuição prática foi a geração e disponibilização de um *corpus* inédito, contendo 192.900 documentos e

mais de **10,4 milhões de entidades anotadas** (ver Seção 4.1.6). Conforme demonstrado no comparativo com trabalhos relacionados (ver Tabela 3.1), este volume de dados supera, em ordens de magnitude, os recursos disponíveis anteriormente na literatura brasileira, mitigando o histórico gargalo no treinamento de modelos supervisionados neste domínio;

2. **Validação Científica:** A metodologia e os resultados preliminares desta pesquisa foram validados pelos pares por meio da publicação e da apresentação no **Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2024)** [15]. Isso confere respaldo técnico às decisões de *design* dos *ensembles* aqui propostos;

3. **Impacto no Projeto *Deep Vacuity*:** O componente de REN desenvolvido oferece ao projeto parceiro uma ferramenta capaz de estruturar dados históricos de convênios, potencializando as ferramentas de fiscalização e de transparência ativa.

## 6.3 Limitações do Trabalho

Apesar dos avanços, é necessário reconhecer as limitações inerentes à abordagem adotada, que devem ser consideradas na interpretação dos resultados:

- **Dependência de Anotação Automática:** A anotação do *corpus* foi realizada via regras heurísticas, resultando no que a literatura denomina *Silver Standard* (padrão prata). Este tipo de recurso caracteriza-se por um grande volume de dados, embora apresente ruídos inerentes ao processo automático. Tal abordagem distingue-se do *Gold Standard* (padrão-ouro), cuja validação manual por humanos o torna a referência livre de erros. Consequentemente, existe o risco dos modelos terem aprendido a replicar certos vieses da anotação automática;

- ***Trade-off* de Custo Computacional (Inferência vs. Treinamento):** A utilização de um *ensemble* com sete modelos implica um custo computacional linearmente maior na fase de predição (aproximadamente 7x mais lento do que um modelo único). Contudo, conforme detalhado na análise de custo (Seção 5.5), esta arquitetura apresenta-se como uma alternativa mais viável do que a busca exaustiva de hiperparâmetros ou o pré-treinamento de modelos do zero. Enquanto essas abordagens demandariam recursos de treinamento proibitivos (estimados em mais de 60x o custo base) com ganhos incertos, o *ensemble* garantiu robustez imediata. Assim, a latência de inferência é assumida como um custo aceitável para o processamento em lote (*batch*) de documentos históricos, como os de convênios no DOU, em detrimento da complexidade de treinamento;

- **Desbalanceamento de Classes:** Apesar da escala massiva do *corpus*, algumas categorias de entidades permaneceram sub-representadas, o que resultou em desempenho nulo para classes raras.

## 6.4 Trabalhos Futuros

A partir das lacunas identificadas e dos resultados obtidos, sugerem-se as seguintes direções para a continuidade desta pesquisa:

1. **Criação de um *"Gold Standard"*:** Recomenda-se a anotação manual de um subconjunto representativo dos dados. A criação desse gabarito revisado por humanos (*Gold Standard*) permitirá avaliar com maior precisão o quanto os modelos estão aprendendo a generalização linguística real versus apenas replicando as regras da anotação automática;
2. **Avaliação de Grandes Modelos de Linguagem (LLMs):** Investigar o desempenho de LLMs generativos modernos (como a família GPT ou Llama) na extração *few-shot* (com poucos exemplos). Estudos recentes sugerem que, embora LLMs tenham alto custo, eles podem servir como anotadores auxiliares ou complementos aos modelos supervisionados em cenários de poucos dados [136];
3. **Otimização e Compressão de Modelos:** Para mitigar o custo computacional do *ensemble*, sugere-se explorar técnicas de compressão de modelos, como *pruning* (poda de parâmetros redundantes da rede neural) ou *knowledge distillation* (destilação de conhecimento), visando manter a acurácia da combinação de modelos com uma maior eficiência computacional [90];
4. **Tratamento de Classes Raras:** Investigar técnicas específicas para lidar com o desbalanceamento extremo, como a geração de dados sintéticos para as entidades com baixo desempenho.

Em suma, esta dissertação demonstra que a combinação de modelos robustos e de dados em larga escala é um caminho viável e eficaz para iluminar o vasto acervo de documentos públicos brasileiros, transformando a burocracia digital em inteligência governamental.

# Referências


[1] Li, Jing, Aixin Sun, Jianglei Han e Chenliang Li: *A survey on deep learning for named entity recognition : Extended abstract*. Em *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, páginas 3817–3818, 2023. xi, 8, 9, 10, 11, 12, 13, 14, 16, 17, 22

[2] Face, Hugging: *The hugging face course, 2022*. `https://huggingface.co/course`, 2022. [Online; accessed <today>]. xi, 20, 21, 22, 23

[3] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser e Illia Polosukhin: *Attention is all you need*. Advances in neural information processing systems, 30, 2017. xi, 1, 3, 19, 20, 24, 25, 26

[4] Peres, Vinícius Araújo: *Classificação de entidades textuais nomeadas em publicações de diários oficiais utilizando transformers*, 2023. `https://bdm.unb.br/handle/10483/39229`, Trabalho de conclusão de curso (Bacharelado em Engenharia Mecatrônica), Universidade de Brasília, Brasília, 2023. xi, xiii, 3, 36, 39, 41, 42, 43, 44, 46

[5] Tay, Yi, Mostafa Dehghani, Dara Bahri e Donald Metzler: *Efficient transformers: A survey*, dezembro 2022. `https://doi.org/10.1145/3530811`. 1

[6] Santos, Diana, Nuno Seco, Nuno Cardoso e Rui Vilela: *Harem: An advanced ner evaluation contest for portuguese*. Em *quot; In Nicoletta Calzolari; Khalid Choukri; Aldo Gangemi; Bente Maegaard; Joseph Mariani; Jan Odjik; Daniel Tapias (ed) Proceedings of the 5 th International Conference on Language Resources and Evaluation (LREC'2006)(Genoa Italy 22-28 May 2006)*, 2006. 1

[7] Li, Jing, Aixin Sun, Jianglei Han e Chenliang Li: *A survey on deep learning for named entity recognition*, janeiro 2022. `https://doi.org/10.1109/tkde.2020.2981314`. 1, 17

[8] Rodríguez, Marcia Marina e Byron Leite Dantas Bezerra: *Processamento de linguagem natural para reconhecimento de entidades nomeadas em textos jurídicos de atos administrativos (portarias)*, abril 2020. `https://doi.org/10.25286/repa.v5i1.1204`. 1

[9] Brazil, Brasil: *Imprensa nacional*. 2025. `http://www.in.gov.br`, visited: 2022-05-01. 1, 40

[10] Possamai, Ana Júlia e Vitória Gonzatti de Souza: *Transparência e dados abertos governamentais: Possibilidades e desafios a partir da lei de acesso À informação*, janeiro 2020. https://doi.org/10.21118/apgs.v12i2.5872. 1

[11] *Decreto nº 11.531, de 16 de maio de 2023.* https://proplad.ufu.br/legislacoes/decreto-no-11531-de-16-de-maio-de-2023, acesso em 2024-08-20. 1, 5

[12] Carvalho, Leonardo Rebouças de, Felipe LS Mendes, Jefferson Chaves, Marcos C Lima, Flavio Elias Gomes de Deus, Aletéia PF Araújo e Flavio de Barros Vidal: *Deep-vacuity: A proposal of a machine learning platform based on high-performance computing architecture for insights on government of brazil official gazettes.* Em *WEBIST*, páginas 136–143, 2022. 1, 79

[13] Sagi, Omer e Lior Rokach: *Ensemble learning: A survey.* Wiley-Blackwell, 8(4), fevereiro 2018. https://doi.org/10.1002/widm.1249. 3, 26

[14] Sagi, Omer e Lior Rokach: *Ensemble learning: A survey.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8, 2018. https://api.semanticscholar.org/CorpusID:49291826. 3, 28, 32

[15] Sirqueira, Eutino e Flávio Vidal: *Evaluation of named entity recognition using ensemble in transformers models for brazilian public texts.* Em *Anais do XXI Encontro Nacional de Inteligência Artificial e Computacional*, páginas 966–977, Porto Alegre, RS, Brasil, 2024. SBC. https://sol.sbc.org.br/index.php/eniac/article/view/33859. 3, 81

[16] Brasil: *Lei n. 8.666, de 21 de junho de 1993.* 1993. http://www.planalto.gov.br/ccivil_03/leis/l8666cons.htm, Accessed: 2021-11-06. 5

[17] *LEI nº 14.133, DE 1º DE ABRIL DE 2021.* https://proplad.ufu.br/legislacoes/lei-no-14133-de-1o-de-abril-de-2021-0, acesso em 2024-08-20. 5

[18] Di Pietro, Maria Sylvia Zanella: *Direito Administrativo.* Forense, Rio de Janeiro, 35ª edição, 2022. 6

[19] Meirelles, Hely Lopes: *Direito Administrativo Brasileiro.* Malheiros, São Paulo, 42ª edição, 2016. 6

[20] Brazil: *Portal da transparência.* https://portaldatransparencia.gov.br/pagina-interna/603415-dicionario-de-dados-convenios, 2022. visited: 2022-01-23. 6, 36

[21] Jurafsky, Daniel e James H. Martin: *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models.* 3rd edição, 2025. https://web.stanford.edu/~jurafsky/slp3/, Online manuscript released August 24, 2025. 7

[22] LeCun, Yann, Yoshua Bengio e Geoffrey Hinton: *Deep learning.* nature, 521(7553):436–444, 2015. 7, 8, 17

[23] Young, Tom, Devamanyu Hazarika, Soujanya Poria e Erik Cambria: *Recent trends in deep learning based natural language processing*. ieee Computational intelligenCe magazine, 13(3):55–75, 2018. 7, 8

[24] Roy, Aurko, Mohammad Saffar, Ashish Vaswani e David Grangier: *Efficient content-based sparse attention with routing transformers*. Transactions of the Association for Computational Linguistics, 9:53–68, 2021. 7

[25] Chiche, Alebachew e Betselot Yitagesu: *Part of speech tagging: a systematic review of deep learning and machine learning approaches*. Journal of Big Data, 9(1):10, 2022. 7

[26] Wang, Xinyu, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang e Kewei Tu: *Automated concatenation of embeddings for structured prediction*. arXiv preprint arXiv:2010.05006, 2020. 7, 32

[27] Li, Jiaqi, Ming Liu, Bing Qin e Ting Liu: *A survey of discourse parsing*. Frontiers of Computer Science, 16(5):165329, 2022. 7

[28] Peters, Matthew E, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee e Luke Zettlemoyer: *Deep contextualized word representations. arxiv 2018*. arXiv preprint arXiv:1802.05365, 12, 2018. 8

[29] Khan, Jawad, Niaz Ahmad, Shah Khalid, Farman Ali e Youngmoon Lee: *Sentiment and context-aware hybrid dnn with attention for text sentiment classification*. IEEE Access, 11:28162–28179, 2023. 8

[30] Awasthi, Ishitva, Kuntal Gupta, Prabjot Singh Bhogal, Sahejpreet Singh Anand e Piyush Kumar Soni: *Natural language processing (nlp) based text summarization-a survey*. Em *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, páginas 1310–1317. IEEE, 2021. 8

[31] Wang, Haifeng, Hua Wu, Zhongjun He, Liang Huang e Kenneth Ward Church: *Progress in machine translation*. Engineering, 18:143–153, 2022. 8

[32] Biancofiore, Giovanni Maria, Yashar Deldjoo, Tommaso Di Noia, Eugenio Di Sciascio e Fedelucio Narducci: *Interactive question answering systems: Literature review*. ACM Computing Surveys, 56(9):1–38, 2024. 8

[33] Wang, Jiajia, Jimmy Xiangji Huang, Xinhui Tu, Junmei Wang, Angela Jennifer Huang, Md Tahmid Rahman Laskar e Amran Bhuiyan: *Utilizing bert for information retrieval: Survey, applications, resources, and challenges*. ACM Computing Surveys, 56(7):1–33, 2024. 8

[34] Eftimov, Tome, Barbara Koroušić Seljak e Peter Korošec: *A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations*. PloS one, 12(6):e0179488, 2017. 8, 14

[35] Keraghel, Imed, Stanislas Morbieu e Mohamed Nadif: *Recent advances in named entity recognition: A comprehensive survey and comparative study*, 2024. `https://arxiv.org/abs/2401.10825`. 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19

[36] Grishman, Ralph e Beth Sundheim: *Message understanding conference-6: a brief history.* Em *Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, COLING '96, página 466–471, USA, 1996. Association for Computational Linguistics. https://doi.org/10.3115/992628.992709. 8, 12

[37] Nadeau, David e Satoshi Sekine: *A survey of named entity recognition and classification.* Lingvisticae Investigationes, 30, agosto 2007. 8, 12, 15

[38] Yadav, Vikas e Steven Bethard: *A survey on recent advances in named entity recognition from deep learning models.* arXiv preprint arXiv:1910.11470, 2019. 8, 11, 12

[39] Huang, Zhiheng, Wei Xu e Kai Yu: *Bidirectional lstm-crf models for sequence tagging. arxiv 2015.* arXiv preprint arXiv:1508.01991, 2015. 9, 19

[40] Honnibal, Matthew, Ines Montani, Sofie Van Landeghem e Adriane Boyd: *spaCy: Industrial-strength natural language processing in python.* 2020. 10, 12, 42, 44, 49, 51, 57

[41] Sufi, Fahim K, Imran Razzak e Ibrahim Khalil: *Tracking anti-vax social movement using ai-based social media monitoring.* IEEE Transactions on Technology and Society, 3(4):290–299, 2022. 10

[42] Ji, Yuanze, Bobo Li, Jun Zhou, Fei Li, Chong Teng e Donghong Ji: *Cmner: A chinese multimodal ner dataset based on social media.* arXiv preprint arXiv:2402.13693, 2024. 10

[43] Bunescu, Razvan e Marius Pasca: *Using encyclopedic knowledge for named entity disambiguation.* Em *11th conference of the European Chapter of the Association for Computational Linguistics*, páginas 9–16, 2006. 10

[44] Dredze, Mark, Paul McNamee, Delip Rao, Adam Gerber, Tim Finin *et al.*: *Entity disambiguation for knowledge base population.* Em *Proceedings of the 23rd international conference on computational linguistics*, páginas 277–285, 2010. 10

[45] Al-Qawasmeh, Omar, Mohammad Al-Smadi e Nisreen Fraihat: *Arabic named entity disambiguation using linked open data.* Em *2016 7th International Conference on Information and Communication Systems (ICICS)*, páginas 333–338. IEEE, 2016. 10

[46] Mollá, Diego, Menno Van Zaanen e Daniel Smith: *Named entity recognition for question answering.* Em *Australasian Language Technology Association Workshop*, páginas 51–58. Australasian Language Technology Association, 2006. 11

[47] Zhang, Yuanzhe, Kang Liu, Shizhu He, Guoliang Ji, Zhanyi Liu, Hua Wu e Jun Zhao: *Question answering over knowledge base with neural attention combining global knowledge information.* arXiv preprint arXiv:1606.00979, 2016. 11

[48] Hkiri, Emna, Souheyl Mallat e Mounir Zrigui: *Arabic-english text translation leveraging hybrid ner.* Em *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, páginas 124–131, 2017. 11

[49] Li, Zhen, Dan Qu, Chaojie Xie, Wenlin Zhang e Yanxia Li: *Language model pre-training method in machine translation based on named entity recognition*. International Journal on Artificial Intelligence Tools, 29(07n08):2040021, 2020. 11

[50] Sang, Erik Tjong Kim: *Introduction to the conll-2002 shared task: Language-independent named entity recognition*. ArXiv, cs.CL/0209010, 2002. 11

[51] Sang, Erik Tjong Kim e Fien De Meulder: *Introduction to the conll-2003 shared task: Language-independent named entity recognition*. Em *CoNLL*, 2003. 11, 12, 31

[52] Walker, Christopher, Stephanie Strassel, Julie Medero e Kazuaki Maeda: *Ace 2005 multilingual training corpus*. Linguistic Data Consortium, Philadelphia, 57:45, 2006. 12

[53] Manning, Christopher D., Prabhakar Raghavan e Hinrich Schütze: *Introduction to Information Retrieval*. Cambridge University Press, 2008. 12, 13, 57

[54] Singh, Umrinderpal, Vishal Goyal e Gurpreet Singh Lehal: *Named entity recognition system for urdu*. Em *International Conference on Computational Linguistics*, 2012. `https://api.semanticscholar.org/CorpusID:17178313`. 14

[55] Goyal, Archana, Vishal Gupta e Manish Kumar: *Recent named entity recognition and classification techniques: A systematic review*. Computer Science Review, 29:21–43, 2018, ISSN 1574-0137. `https://www.sciencedirect.com/science/article/pii/S1574013717302782`. 14, 15, 16

[56] Quimbaya, Alexandra Pomares, Alejandro Sierra Múnera, Rafael Andrés González Rivera, Julián Camilo Daza Rodríguez, Oscar Mauricio Muñoz Velandia, Angel Alberto Garcia Peña e Cyril Labbé: *Named entity recognition over electronic health records through a combined dictionary-based approach*. Procedia Computer Science, 100:55–61, 2016, ISSN 1877-0509. `https://www.sciencedirect.com/science/article/pii/S1877050916322906`, International Conference on ENTERprise Information Systems/International Conference on Project MANagement/International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN / HCist 2016. 14

[57] Riaz, Kashif: *Rule-based named entity recognition in Urdu*. Em Kumaran, A e Haizhou Li (editores): *Proceedings of the 2010 Named Entities Workshop*, páginas 126–135, Uppsala, Sweden, julho 2010. Association for Computational Linguistics. `https://aclanthology.org/W10-2419/`. 15

[58] Wu, Lang Tao, Jia Rui Lin, Shuo Leng, Jiu Lin Li e Zhen Zhong Hu: *Rule-based information extraction for mechanical-electrical-plumbing-specific semantic web*. Automation in Construction, 135:104108, 2022. 15

[59] Mengliev, Davlatyor B, Vladimir B Barakhnin, Mukhammadjon Atakhanov, Bahodir B Ibragimov, Mukhriddin Eshkulov e Bobur Saidov: *Developing rule-based and gazetteer lists for named entity recognition in uzbek language: geographical*

*names*. Em *2023 IEEE XVI International Scientific and Technical Conference Actual Problems of Electronic Instrument Engineering (APEIE)*, páginas 1500–1504. IEEE, 2023. 15

[60] IBM: *What is machine learning?*, 2025. `https://www.ibm.com/think/topics/machine-learning`, Acesso em: 28 ago. 2025. 15, 17, 23

[61] Lample, Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami e Chris Dyer: *Neural architectures for named entity recognition*. Em *NAACL*, 2016. 16, 19

[62] Bikel, Daniel M, Richard Schwartz e Ralph M Weischedel: *An algorithm that learns what's in a name*. Machine learning, 34:211–231, 1999. 16

[63] Morwal, Sudha, Nusrat Jahan e Deepti Chopra: *Named entity recognition using hidden markov model (hmm)*. International Journal on Natural Language Computing (IJNLC) Vol, 1, 2012. 16

[64] Borthwick, Andrew Eliot: *A maximum entropy approach to named entity recognition*. New York University, 1999. 16

[65] Saha, Sujan Kumar, Sudeshna Sarkar e Pabitra Mitra: *Feature selection techniques for maximum entropy based biomedical named entity recognition*. Journal of biomedical informatics, 42(5):905–911, 2009. 16

[66] Isozaki, Hideki e Hideto Kazawa: *Efficient support vector classifiers for named entity recognition*. Em *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002. 16

[67] Saha, Sujan Kumar, Shashi Narayan, Sudeshna Sarkar e Pabitra Mitra: *A composite kernel for named entity recognition*. Pattern Recognition Letters, 31(12):1591–1597, 2010. 16

[68] McCallum, Andrew e Wei Li: *Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons*. 2003. 16

[69] Settles, Burr: *Biomedical named entity recognition using conditional random fields and rich feature sets*. Em *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (NLPBA/BioNLP)*, páginas 107–110, 2004. 16

[70] Lafferty, John D., Andrew McCallum e Fernando C.N. Pereira: *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. Em *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, 2001. 16

[71] Lample, Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami e Chris Dyer: *Neural architectures for named entity recognition*. arXiv preprint arXiv:1603.01360, 2016. 16, 18, 19

[72] Thenmalar, S, J Balaji e TV Geetha: *Semi-supervised bootstrapping approach for named entity recognition.* arXiv preprint arXiv:1511.06833, 2015. 16

[73] Van Engelen, Jesper E e Holger H Hoos: *A survey on semi-supervised learning.* Machine Learning, 109(2):373–440, 2020. 17

[74] Norvig, P. e S. Russell: *Inteligência artificial: Tradução da 3a Edição.* Elsevier Brasil, 2014, ISBN 9788535251418. `https://books.google.com.br/books?id=BsNeAwAAQBAJ`. 17

[75] Shinyama, Yusuke e Satoshi Sekine: *Named entity discovery using comparable news articles.* Em *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, páginas 848–853, Geneva, Switzerland, aug 23–aug 27 2004. COLING. `https://aclanthology.org/C04-1122/`. 17

[76] Bonnefoy, Ludovic, Patrice Bellot e Michel Benoit: *An unsupervised measure of the degree of belonging of an entity to a type.* Traitement Automatique des Langues Naturelles, página 75, 2011. 17

[77] Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado e Jeff Dean: *Distributed representations of words and phrases and their compositionality.* Advances in neural information processing systems, 26, 2013. 18

[78] Pennington, Jeffrey, Richard Socher e Christopher D Manning: *Glove: Global vectors for word representation.* Em *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, páginas 1532–1543, 2014. 18

[79] Bojanowski, Piotr, Edouard Grave, Armand Joulin e Tomas Mikolov: *Enriching word vectors with subword information.* Transactions of the association for computational linguistics, 5:135–146, 2017. 18

[80] Wang, Yu, Bin Xia, Zheng Liu, Yun Li e Tao Li: *Named entity recognition for chinese telecommunications field based on char2vec and bi-lstms.* Em *2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, páginas 1–7. IEEE, 2017. 18

[81] Lecun, Y., L. Bottou, Y. Bengio e P. Haffner: *Gradient-based learning applied to document recognition.* Proceedings of the IEEE, 86(11):2278–2324, 1998. 18

[82] Ma, Xuezhe e Eduard Hovy: *End-to-end sequence labeling via bi-directional lstm-cnns-crf.* arXiv preprint arXiv:1603.01354, 2016. 18, 19

[83] Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu e Pavel Kuksa: *Natural language processing (almost) from scratch.* Journal of machine learning research, 12:2493–2537, 2011. 18, 19

[84] Hochreiter, Sepp e Jürgen Schmidhuber: *Long Short-Term Memory.* Neural Computation, 9(8):1735–1780, novembro 1997, ISSN 0899-7667. `https://doi.org/10.1162/neco.1997.9.8.1735`. 19

[85] Banik, Nayan e Md Hasan Hafizur Rahman: *Gru based named entity recognition system for bangla online newspapers*. Em *2018 International Conference on Innovation in Engineering and Technology (ICIET)*, páginas 1–6. IEEE, 2018. 19

[86] Forney, Jr., G. David: *The viterbi algorithm*. Proceedings of the IEEE, 61(3):268–278, 1973. 19

[87] Radford, Alec e Karthik Narasimhan: *Improving language understanding by generative pre-training*. 2018. `https://api.semanticscholar.org/CorpusID:49313245`. 20, 21, 26

[88] Devlin, Jacob: *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805, 2018. 20, 22, 26, 49, 57

[89] Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever *et al.*: *Language models are unsupervised multitask learners*. OpenAI blog, 1(8):9, 2019. 20

[90] Sanh, V: *Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter*. arXiv preprint arXiv:1910.01108, 2019. 20, 82

[91] Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov e Luke Zettlemoyer: *Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*. arXiv preprint arXiv:1910.13461, 2019. 20, 49, 57

[92] Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li e Peter J Liu: *Exploring the limits of transfer learning with a unified text-to-text transformer*. Journal of machine learning research, 21(140):1–67, 2020. 20

[93] Mann, Ben, Nick Ryder, Melanie Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal *et al.*: *Language models are few-shot learners*. arXiv preprint arXiv:2005.14165, 1(3):3, 2020. 20

[94] Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray *et al.*: *Training language models to follow instructions with human feedback*. Advances in neural information processing systems, 35:27730–27744, 2022. 20

[95] Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar *et al.*: *Llama: Open and efficient foundation language models*. arXiv preprint arXiv:2302.13971, 2023. 20

[96] Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix e William El Sayed: *Mistral 7b*, 2023. `https://arxiv.org/abs/2310.06825`. 20

[97] Team, Gemma, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé *et al.*: *Gemma 2: Improving open language models at a practical size*. arXiv preprint arXiv:2408.00118, 2024. 20

[98] Allal, Loubna Ben, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav *et al.*: *Smollm2: When smol goes big–data-centric training of a small language model*. arXiv preprint arXiv:2502.02737, 2025. 20

[99] Wolf, Thomas: *Transformers: State-of-the-art natural language processing*. arXiv preprint arXiv:1910.03771, 2020. 22

[100] Khan, Azal Ahmad, Omkar Chaudhari e Rohitash Chandra: *A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation*, junho 2024. `https://doi.org/10.1016/j.eswa.2023.122778`. 26

[101] Hastie, Trevor, Robert Tibshirani e Jerome Friedman: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2009. 26, 27, 29

[102] Geman, Stuart, Elie Bienenstock e René Doursat: *Neural networks and the bias/variance dilemma*. Neural computation, 4(1):1–58, 1992. 26, 27

[103] Rokach, Lior: *Ensemble-based classifiers*. Artificial intelligence review, 33:1–39, 2010. 27, 29, 54

[104] Breiman, Leo: *Bagging predictors*. Machine learning, 24:123–140, 1996. 27

[105] Breiman, Leo: *Random forests*. Machine learning, 45:5–32, 2001. 27

[106] Freund, Yoav e Robert E Schapire: *A decision-theoretic generalization of on-line learning and an application to boosting*. Journal of computer and system sciences, 55(1):119–139, 1997. 27

[107] Friedman, Jerome H.: *Greedy function approximation: a gradient boosting machine*. Annals of statistics, páginas 1189–1232, 2001. 27

[108] Wolpert, David H: *Stacked generalization*. Neural networks, 5(2):241–259, 1992. 28

[109] Zhang, Hongzhi e M Omair Shafiq: *Survey of transformers and towards ensemble learning using transformers for natural language processing*. Journal of big Data, 11(1):25, 2024. 29, 32

[110] Singh, Aayushi, Sumit Singh Singh e Uma Shanker Tiwary: *Enhancing hindi named entity recognition through ensemble learning*, dezembro 2023. `https://ieeexplore.ieee.org/document/10434322/`. 29, 33

[111] Rouhizadeh, Hossein e Douglas Teodoro: *Ds4dh at semeval-2022 task 11: Multilingual named entity recognition using an ensemble of transformer-based language models*. Em *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, páginas 1543–1548, 2022. https://doi.org/10.18653/v1/2022.semeval-1.212. 30, 33, 34

[112] Zheng, Jian-Yu e Jin Sun: *Ensembles of bert models for ancient chinese processing*, setembro 2023. 30, 33, 34

[113] Sun, Jian, Rongchuan Tang, Lu Xiang, Feifei Zhai e Yu Zhou: *Multi-strategy fusion for medical named entity recognition*, dezembro 2021. https://doi.org/10.1109/ialp54817.2021.9675228. 30, 33, 34

[114] Schweter, Stefan e Alan Akbik: *Flert: Document-level features for named entity recognition*. arXiv preprint arXiv:2011.06993, 2020. 32

[115] Yamada, Ikuya, Akari Asai, Hiroyuki Shindo, Hideaki Takeda e Yuji Matsumoto: *Luke: deep contextualized entity representations with entity-aware self-attention*. arXiv preprint arXiv:2010.01057, 2020. 32

[116] Zhou, Wenxuan e Muhao Chen: *Learning from noisy labels for entity-centric information extraction*. arXiv preprint arXiv:2104.08656, 2021. 32

[117] Alles, Vanderlei Jandir: *Construção de um corpus para extrair entidades nomeadas do diário oficial da união utilizando aprendizado supervisionado*. 2018. 32

[118] Alles, Vanderlei J., William F. Giozza e Robson de Oliveira Alburquerque: *Natural language processing to classify named entities of the brazilian union official diary*, junho 2018. https://ieeexplore.ieee.org/abstract/document/8399215. 32

[119] Araujo, Pedro Henrique Luz de, Teófilo de Campos, Renato Resende Ribeiro de Oliveira, Matheus Stauffer, Samuel Couto e Paulo Henrique de Souza Bermejo: *Lener-br: A dataset for named entity recognition in brazilian legal text*, janeiro 2018. 32, 34

[120] Silva, Messias Gomes da: *Reconhecimento de entidades nomeadas em documentos de editais de compras utilizando aprendizado profundo*, janeiro 2022. https://repositorio.ifes.edu.br/handle/123456789/2944. 33, 34

[121] Souza, Fábio, Rodrigo Nogueira e Roberto Lotufo: *BERTimbau: pretrained BERT models for Brazilian Portuguese*. Em *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*, 2020. 33, 49, 57

[122] Mascarenhas, Ana, Rita de Lhama e Helena Sousa: *Legalbertpt: a Bert model for the Portuguese legal language*. arXiv preprint arXiv:2211.02102, 2022. 33

[123] Albanaz, Jennifer Olivia Leiria: *Reconhecimento de entidades nomeadas em resultados de licitações publicados em diários oficiais*, janeiro 2020. https://hdl.handle.net/1884/71065. 33

[124] Pietiläinen, Aapo e Shaoxiong Ji: *Aaltonlp at semeval-2022 task 11: Ensembling task-adaptive pretrained transformers for multilingual complex ner*. Em *International Workshop on Semantic Evaluation*, páginas 1477–1482. The Association for Computational Linguistics, 2022. 33

[125] Aurpa, Tanjim Taharat e Md Shoaib Ahmed: *An ensemble novel architecture for bangla mathematical entity recognition (mer) using transformer based learning*. Heliyon, 10(3), 2024. 33

[126] Brazil: *Portal da transparência*. `https://portaldatransparencia.gov.br/download-de-dados/convenios`, 2022. visited: 2022-01-23. 36

[127] Bhardwaj, Bhavya, Syed Ishtiyaq Ahmed, J Jaiharie, R Sorabh Dadhich e M Ganesan: *Web scraping using summarization and named entity recognition (ner)*. Em *2021 7th international conference on advanced computing and communication systems (ICACCS)*, volume 1, páginas 261–265. IEEE, 2021. 40

[128] Mintz, Mike, Steven Bills, Rion Snow e Dan Jurafsky: *Distant supervision for relation extraction without labeled data*. Em *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, páginas 1003–1011, 2009. 46

[129] Lan, Zhenzhong: *Albert: A lite bert for self-supervised learning of language representations*. arXiv preprint arXiv:1909.11942, 2019. 49, 57, 77

[130] Clark, K: *Electra: Pre-training text encoders as discriminators rather than generators*. arXiv preprint arXiv:2003.10555, 2020. 49, 57

[131] Domingues, Maicon: *Language model in the legal domain in portuguese*. `https://huggingface.co/dominguesm/legal-bert-base-cased-ptbr/`, 2022. 49, 57

[132] Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer e Veselin Stoyanov: *Unsupervised cross-lingual representation learning at scale*. arXiv preprint arXiv:1911.02116, 2019. 49, 57

[133] Kingma, Diederik P. e Jimmy Ba: *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980, 2014. 49, 52

[134] Amari, Shun ichi: *Backpropagation and stochastic gradient descent method*. Neurocomputing, 5(4-5):185–196, 1993. 50

[135] Caseli, H. M. e M. G. V. Nunes (editores): *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, 3ª edição, 2024, ISBN 978-65-01-20581-6. `https://brasileiraspln.com/livro-pln/3a-edicao/`. 61, 65, 68

[136] Ding, Bosheng, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty e Lidong Bing: *Is gpt-3 a good data annotator?* Em *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 11173–11195, 2023. 82

# Apêndice A

# Apêndices

Neste apêndice, são apresentados os resultados brutos dos testes de variação dos parâmetros discriminados por categoria de entidade.

Tabela A.1: Comparativo detalhado: Estratégias de Seleção de Melhor Modelo (Baseado em F1 vs. Revocação).

| **Entidade** | **E1 (Best F1)** | | | **Teste F (Best Recall)** | | |
|---|---|---|---|---|---|---|
| | **P** | **R** | **F** | **P** | **R** | **F** |
| CÓDIGO CONVENENTE | 0,9945 | 0,4561 | 0,6254 | 0,9945 | 0,4562 | 0,6255 |
| CÓDIGO SIAFI MUNICÍPIO | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| CÓDIGO UG CONCEDENTE | 0,7812 | 0,5799 | 0,6657 | 0,7786 | 0,5812 | 0,6656 |
| CÓDIGO ÓRGÃO CONCEDENTE | 0,5794 | 0,5536 | 0,5662 | 0,5358 | 0,5358 | 0,5358 |
| CÓDIGO ÓRGÃO SUPERIOR | 0,2530 | 0,8400 | 0,3889 | 0,5663 | 0,7833 | 0,6573 |
| DATA FINAL VIGÊNCIA | 0,9007 | 0,4660 | 0,6142 | 0,9005 | 0,4663 | 0,6144 |
| DATA INÍCIO VIGÊNCIA | 0,9191 | 0,6286 | 0,7466 | 0,9190 | 0,6287 | 0,7466 |
| DATA ÚLTIMA LIBERAÇÃO | 0,2222 | 0,1791 | 0,1983 | 0,2222 | 0,1791 | 0,1983 |
| NOME CONVENENTE | 0,9401 | 0,3746 | 0,5358 | 0,9328 | 0,3738 | 0,5337 |
| NOME MUNICÍPIO | 0,7996 | 0,3796 | 0,5148 | 0,8042 | 0,3792 | 0,5154 |
| NOME UG CONCEDENTE | 0,5381 | 0,3423 | 0,4185 | 0,5678 | 0,1215 | 0,2001 |
| NOME ÓRGÃO CONCEDENTE | 0,9802 | 0,5150 | 0,6752 | 0,9782 | 0,5509 | 0,7049 |
| NOME ÓRGÃO SUPERIOR | 0,9993 | 0,8432 | 0,9146 | 0,9978 | 0,8418 | 0,9132 |
| NÚMERO CONVÊNIO | 0,9990 | 0,5228 | 0,6864 | 0,9996 | 0,5212 | 0,6851 |
| NÚMERO ORIGINAL | 0,8297 | 0,1839 | 0,3011 | 0,8319 | 0,1843 | 0,3017 |
| NÚMERO PROCESSO | 0,5177 | 0,6234 | 0,5657 | 0,5166 | 0,6284 | 0,5670 |
| OBJETO DO CONVÊNIO | 0,4544 | 0,4179 | 0,4354 | 0,4542 | 0,4140 | 0,4332 |
| SITUAÇÃO CONVÊNIO | 0,0000 | 0,0000 | 0,0000 | 0,1304 | 0,2143 | 0,1622 |
| TIPO CONVENENTE | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| TIPO ENTE CONVENENTE | 0,9715 | 0,9228 | 0,9465 | 0,9712 | 0,9228 | 0,9464 |
| TIPO INSTRUMENTO | 0,9846 | 0,6575 | 0,7885 | 0,9846 | 0,6553 | 0,7869 |
| UF | 0,9877 | 0,5736 | 0,7258 | 0,9871 | 0,5717 | 0,7240 |
| VALOR CONTRAPARTIDA | 0,9293 | 0,2883 | 0,4401 | 0,9282 | 0,2881 | 0,4397 |
| VALOR CONVÊNIO | 0,6945 | 0,3347 | 0,4517 | 0,6790 | 0,3332 | 0,4470 |
| VALOR LIBERADO | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| VALOR ÚLTIMA LIBERAÇÃO | 0,0952 | 0,1141 | 0,1038 | 0,1038 | 0,1166 | 0,1098 |

| Entidade | E2 (Limiar 4) | | | Teste A (Limiar 3) | | | Teste B (Limiar 6) | | |
|---|---|---|---|---|---|---|---|---|---|
| | **P** | **R** | **F** | **P** | **R** | **F** | **P** | **R** | **F** |
| CÓDIGO CONVENENTE | 0,9705 | 0,6195 | 0,7563 | 0,9878 | 0,5591 | 0,7141 | 0,9573 | 0,6238 | 0,7554 |
| CÓDIGO SIAFI MUN. | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| CÓDIGO UG CONCED. | 0,6103 | 0,8089 | 0,6957 | 0,7260 | 0,7448 | 0,7353 | 0,4372 | 0,8117 | 0,5683 |
| CÓDIGO ÓRGÃO CONC. | 0,4081 | 0,9562 | 0,5721 | 0,4486 | 0,9290 | 0,6050 | 0,3894 | 0,9615 | 0,5543 |
| CÓDIGO ÓRGÃO SUP. | 0,0602 | 1,0000 | 0,1136 | 0,1325 | 1,0000 | 0,2340 | 0,0241 | 1,0000 | 0,0471 |
| DATA FINAL VIGÊNCIA | 0,6181 | 0,6838 | 0,6493 | 0,8226 | 0,5391 | 0,6514 | 0,5255 | 0,7303 | 0,6112 |
| DATA INÍCIO VIGÊNCIA | 0,7783 | 0,7651 | 0,7716 | 0,8610 | 0,6981 | 0,7710 | 0,7436 | 0,8113 | 0,7760 |
| DATA ÚLT. LIBERAÇÃO | 0,1296 | 0,7000 | 0,2188 | 0,2037 | 0,4783 | 0,2857 | 0,1296 | 0,7000 | 0,2188 |
| NOME CONVENENTE | 0,9116 | 0,4722 | 0,6222 | 0,9430 | 0,4385 | 0,5987 | 0,8712 | 0,5351 | 0,6630 |
| NOME MUNICÍPIO | 0,6447 | 0,5246 | 0,5784 | 0,7678 | 0,4415 | 0,5606 | 0,5473 | 0,5811 | 0,5637 |
| NOME UG CONCED. | 0,4661 | 0,8333 | 0,5978 | 0,5042 | 0,1413 | 0,2208 | 0,3475 | 0,9213 | 0,5046 |
| NOME ÓRGÃO CONC. | 0,9185 | 0,8835 | 0,9006 | 0,9636 | 0,7749 | 0,8590 | 0,8970 | 0,9097 | 0,9034 |
| NOME ÓRGÃO SUP. | 0,9982 | 0,8874 | 0,9396 | 0,9990 | 0,8634 | 0,9263 | 0,9967 | 0,8986 | 0,9451 |
| NÚMERO CONVÊNIO | 0,9936 | 0,5609 | 0,7170 | 0,9977 | 0,5432 | 0,7034 | 0,9898 | 0,5665 | 0,7206 |
| NÚMERO ORIGINAL | 0,5752 | 0,3598 | 0,4427 | 0,7828 | 0,2538 | 0,3834 | 0,4641 | 0,4147 | 0,4380 |
| NÚMERO PROCESSO | 0,6277 | 0,7381 | 0,6784 | 0,8135 | 0,6735 | 0,7369 | 0,5090 | 0,7600 | 0,6097 |
| OBJETO DO CONVÊNIO | 0,1871 | 0,6312 | 0,2886 | 0,3760 | 0,5021 | 0,4300 | 0,1352 | 0,7046 | 0,2269 |
| SITUAÇÃO CONVÊNIO | 0,0000 | 0,0000 | 0,0000 | 0,0870 | 1,0000 | 0,1600 | 0,0000 | 0,0000 | 0,0000 |
| TIPO CONVENENTE | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| TIPO ENTE CONVENENTE | 0,9584 | 0,9433 | 0,9508 | 0,9600 | 0,9356 | 0,9477 | 0,9521 | 0,9460 | 0,9491 |
| TIPO INSTRUMENTO | 0,9179 | 0,8249 | 0,8689 | 0,9692 | 0,6974 | 0,8112 | 0,8615 | 0,8358 | 0,8485 |
| UF | 0,9358 | 0,6444 | 0,7633 | 0,9679 | 0,6203 | 0,7561 | 0,9041 | 0,6537 | 0,7588 |
| VALOR CONTRAPARTIDA | 0,6826 | 0,6179 | 0,6486 | 0,8164 | 0,4512 | 0,5812 | 0,6182 | 0,6835 | 0,6493 |
| VALOR CONVÊNIO | 0,2392 | 0,4522 | 0,3129 | 0,4855 | 0,3861 | 0,4301 | 0,1560 | 0,5029 | 0,2382 |
| VALOR LIBERADO | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| VALOR ÚLT. LIBERAÇÃO | 0,3091 | 0,5471 | 0,3950 | 0,4317 | 0,3489 | 0,3859 | 0,1757 | 0,6249 | 0,2743 |

Tabela A.2: Comparativo detalhado: Estratégias de Votação por Maioria (Variação do limiar de votos).

Tabela A.3: Comparativo detalhado: Estratégias de Votação Ponderada (Variação de limiar de probabilidade e pesos).

| Entidade | E3 (Limiar 0.8) | | | Teste C (Limiar 0.3) | | | Teste D (Limiar 0.5) | | | Teste E (Peso Rec) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F | P | R | F | P | R | F | P | R | F |
| CÓDIGO CONVENENTE | 0,9260 | 0,6363 | 0,7543 | 0,9802 | 0,5874 | 0,7346 | 0,9705 | 0,6195 | 0,7563 | 0,9705 | 0,6195 | 0,7563 |
| CÓDIGO SIAFI MUN. | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| CÓDIGO UG CONC. | 0,2740 | 0,8364 | 0,4127 | 0,6771 | 0,7889 | 0,7287 | 0,6103 | 0,8089 | 0,6957 | 0,6103 | 0,8089 | 0,6957 |
| CÓDIGO ÓRGÃO CONC. | 0,3458 | 0,9737 | 0,5103 | 0,4112 | 0,9496 | 0,5739 | 0,4081 | 0,9562 | 0,5721 | 0,4081 | 0,9562 | 0,5721 |
| CÓDIGO ÓRGÃO SUP. | 0,0000 | 0,0000 | 0,0000 | 0,0602 | 1,0000 | 0,1136 | 0,0241 | 1,0000 | 0,0471 | 0,0241 | 1,0000 | 0,0471 |
| DATA FINAL VIGÊNCIA | 0,0000 | 0,0000 | 0,0000 | 0,7225 | 0,6234 | 0,6693 | 0,5255 | 0,7303 | 0,6112 | 0,5255 | 0,7303 | 0,6112 |
| DATA INÍCIO VIGÊNCIA | 0,7002 | 0,8458 | 0,7661 | 0,8146 | 0,7329 | 0,7716 | 0,7783 | 0,7651 | 0,7716 | 0,7783 | 0,7651 | 0,7716 |
| DATA ÚLT. LIBERAÇÃO | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| NOME CONVENENTE | 0,8060 | 0,6186 | 0,7000 | 0,9352 | 0,4504 | 0,6080 | 0,9116 | 0,4722 | 0,6222 | 0,9116 | 0,4722 | 0,6222 |
| NOME MUNICÍPIO | 0,0000 | 0,0000 | 0,0000 | 0,7142 | 0,4825 | 0,5760 | 0,5473 | 0,5811 | 0,5637 | 0,5473 | 0,5811 | 0,5637 |
| NOME UG CONCED. | 0,2669 | 1,0000 | 0,4214 | 0,4958 | 0,8125 | 0,6158 | 0,4661 | 0,8333 | 0,5978 | 0,4661 | 0,8333 | 0,5978 |
| NOME ÓRGÃO CONC. | 0,8662 | 0,9364 | 0,8999 | 0,9467 | 0,8239 | 0,8811 | 0,9185 | 0,8835 | 0,9006 | 0,9185 | 0,8835 | 0,9006 |
| NOME ÓRGÃO SUP. | 0,9944 | 0,9051 | 0,9476 | 0,9987 | 0,8725 | 0,9314 | 0,9982 | 0,8874 | 0,9396 | 0,9982 | 0,8874 | 0,9396 |
| NÚMERO CONVÊNIO | 0,9805 | 0,5715 | 0,7221 | 0,9959 | 0,5517 | 0,7100 | 0,9936 | 0,5609 | 0,7170 | 0,9936 | 0,5609 | 0,7170 |
| NÚMERO ORIGINAL | 0,0000 | 0,0000 | 0,0000 | 0,6674 | 0,3223 | 0,4347 | 0,4319 | 0,4318 | 0,4319 | 0,4352 | 0,4314 | 0,4333 |
| NÚMERO PROCESSO | 0,4602 | 0,7823 | 0,5795 | 0,7786 | 0,6870 | 0,7300 | 0,6277 | 0,7381 | 0,6784 | 0,6277 | 0,7381 | 0,6784 |
| OBJETO DO CONVÊNIO | 0,0000 | 0,0000 | 0,0000 | 0,1352 | 0,7046 | 0,2269 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| SITUAÇÃO CONVÊNIO | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| TIPO CONVENENTE | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| TIPO ENTE CONVENENTE | 0,9441 | 0,9492 | 0,9466 | 0,9617 | 0,9414 | 0,9514 | 0,9584 | 0,9433 | 0,9508 | 0,9584 | 0,9433 | 0,9508 |
| TIPO INSTRUMENTO | 0,7026 | 0,9133 | 0,7942 | 0,9333 | 0,7109 | 0,8071 | 0,9179 | 0,8249 | 0,8689 | 0,9179 | 0,8249 | 0,8689 |
| UF | 0,7202 | 0,7200 | 0,7201 | 0,9592 | 0,6327 | 0,7624 | 0,9358 | 0,6444 | 0,7633 | 0,9355 | 0,6444 | 0,7632 |
| VALOR CONTRAPARTIDA | 0,4191 | 0,7528 | 0,5385 | 0,7459 | 0,5489 | 0,6324 | 0,6826 | 0,6179 | 0,6486 | 0,6826 | 0,6179 | 0,6486 |
| VALOR CONVÊNIO | 0,0473 | 0,6781 | 0,0884 | 0,3531 | 0,4175 | 0,3826 | 0,1560 | 0,5029 | 0,2382 | 0,1560 | 0,5029 | 0,2382 |
| VALOR LIBERADO | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| VALOR ÚLT. LIBERAÇÃO | 0,0000 | 0,0000 | 0,0000 | 0,3787 | 0,4630 | 0,4166 | 0,1757 | 0,6249 | 0,2743 | 0,1756 | 0,6255 | 0,2742 |