Universidade de Brasília
Faculdade de Tecnologia

Classificação de Entidades Textuais Nomeadas em Publicações de Diários Oficiais utilizando Transformers

Vinícius Araújo Peres

PROJETO FINAL DE CURSO

ENGENHARIA DE CONTROLE E AUTOMAÇÃO

Brasília

2023

Universidade de Brasília
Faculdade de Tecnologia

Classificação de Entidades Textuais Nomeadas em Publicações de Diários Oficiais utilizando Transformers

Vinícius Araújo Peres

Projeto Final de Curso submetido como requisito parcial para obtenção do grau de Engenheiro de Controle e Automação

Orientador: Prof. Dr. Flávio de Barros Vidal

Brasília
2023

Peres, Vinícius Araújo.
P437c Classificação de Entidades Textuais Nomeadas em Publicações de Diários Oficiais utilizando Transformers / Vinícius Araújo Peres; orientador Flávio de Barros Vidal. -- Brasília, 2023.
70 p.

Projeto Final de Curso (Engenharia de Controle e Automação) -- Universidade de Brasília, 2023.

1. Processamento de Linguagem Natural. 2. Aprendizado Profundo. 3. Reconhecimento de Entidades Nomeadas. 4. Transformers. I. Vidal, Flávio de Barros, orient. II. Título

Universidade de Brasília
Faculdade de Tecnologia

Classificação de Entidades Textuais Nomeadas em Publicações de Diários Oficiais utilizando Transformers

Vinícius Araújo Peres

Projeto Final de Curso submetido como requisito parcial para obtenção do grau de Engenheiro de Controle e Automação

Trabalho aprovado. Brasília, 20 de Julho de 2023:

Prof. Dr. Flávio de Barros Vidal, UnB/IE/CIC
Orientador

Prof. Dra. Aletéia Araújo, UnB/IE/CIC
Examinador interno

PCF. Dr. Alan de Oliveira Lopes, PF/INC
Examinador externo

PCF. Ms. Marcos Cavalcanti Lima, PF/INC
Examinador externo

Brasília
2023

Este trabalho é dedicado à todas as pessoas que, de alguma forma, contribuíram com a minha jornada acadêmica e profissional, mas principalmente ao meu pai, que fez o possível e o impossível para isso acontecer.

Agradecimentos

Agradeço primeiramente ao meu pai, Valdir Correa Peres, que nunca mediu esforços para possibilitar que eu conseguisse chegar até aqui, algo que ele nunca conseguiu por não ter tido oportunidade. Por isso, essa conquista não é apenas minha, mas também dele.

Agradeço ao Prof. Dr. Flávio de Barros Vidal por toda a orientação, que foi fundamental para a conclusão deste trabalho. Agradeço todo o tempo disponibilizado e também a compreensão nos momentos mais difíceis. Se depender de mim, sua fama de ser um excelente orientador continuará sendo espalhada.

Agradeço também minha namorada, Thais Silva Queiroz, que nesses quase cinco anos juntos, sempre me apoiou e incentivou nos momentos mais difíceis e nunca desistiu de mim, mesmo à centenas de quilômetros de distância.

Por último e não menos importante, agradeço a Universidade de Brasília (UnB) por todas as oportunidades oferecidas e também ao Departamento de Ciência da Computação (CIC) pela disponibilização do "super servidor", que permitiu que este trabalho fosse realizado.

Resumo

O processo de Reconhecimento de Entidades Nomeadas (REN) é uma das atividades fundamentais e formantes da área de Processamento de Linguagem Natural (PLN). Atualmente, existem diversas fontes de informações de dados textuais com valor importante para a sociedade. Dentre estas fontes de informações tem-se as publicações de diários oficiais. Tais publicações possuem diversos elementos textuais de alta relevância, as quais servem tanto para informar a sociedade, quanto para permitir a detecção de suspeita de atividades de conluio e fraudes em contratos e licitações. Assim sendo, esta monografia promove a realização de um fluxo de trabalho utilizando modelos de *Transformer*, de forma a permitir o reconhecimento de tais entidades nos textos de publicações de diários oficiais. A partir da informação de convênios públicos foi construída uma base de publicações anotadas que permite o treinamento de dois modelos de *Transformer*, BERT e RoBERTa, e posteriormente a comparação entre eles. Os resultados obtidos revelaram que ambos os modelos apresentaram desempenho geral bastante semelhante, alcançando F1-Scores de 94% e 94,2%, respectivamente. Porém, com variações no desempenho por tipo de entidade.

Palavras-chave: Processamento de Linguagem Natural. Aprendizado Profundo. Reconhecimento de Entidades Nomeadas. Transformers.

Abstract

The Named Entity Recognition (NER) process is one of the fundamental and formative activities in natural language processing. Currently, various sources of textual data information have significant value to society. Among these sources of information are official gazette publications. Such publications contain several text elements of high relevance that both inform society and enable the detection of suspicion of collusion and fraud activities in contracts and bids. Therefore, this monography promotes the realization of a workflow using Transformer models to allow the recognition of such entities in texts from official gazette publications. Based on public agreement information, an annotated publication database was built to train two Transformer models, BERT and RoBERTa, and, subsequently, compare them. The results revealed that both models showed a very similar overall performance, achieving F1-Scores of 94% and 94,2%, respectively, with variations in performance depending on the entity type.

Keywords: Natural Language Processing. Deep Learning. Named Entity Recognition. Transformers.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

1.1 Motivações e Justificativas

O grande volume financeiro envolvido e a ausência de punições severas são alguns dos motivos que fazem com que licitações públicas sejam alvos de ações fraudulentas, principalmente em obras de infraestrutura pública, nas quais os investimentos são altos, como descreve Santos Nakamura (2018). Esse investimento elevado do governo brasileiro em licitações públicas pode ser exemplificado pelos últimos cinco anos, em que o mesmo desembolsou cerca de R$ 66,6 bilhões em transferências voluntárias, de acordo com MGISP (2023).

Com o advento da pandemia de Covid-19 em 2020 no Brasil e no mundo, as fraudes em processos licitatórios se tornaram mais comuns, principalmente devido à flexibilização das regras para licitações e contratos por conta do estado de calamidade pública, como pode ser visto em Brasil (2020), o que motivou mais ainda a necessidade de se combater as fraudes e recuperar os prejuízos aos cofres públicos. Como pode ser visto no Portal da Transparência (2023a), em 2020 a União investiu cerca de R$ 42,53 bilhões no enfrentamento da Covid-19 e estima-se que, de abril de 2020 a abril de 2022, o prejuízo potencial seja de cerca de R$ 309,2 milhões, como consta em CGU (2023). Dentre os tipos de fraudes mais comuns nesse período, estão a contratação de empresas sem capacidade técnica, conluio entre empresas, sobrepreço e superfaturamento, como investigado por Kleberson Souza (2023).

O processo investigatório de fraudes em licitações e contratos públicos é baseado, entre outros aspectos, no conhecimento adquirido ao longo dos anos de experiência, como observado por Marcos Cavalcanti Lima (2021). Alguns tipos de fraudes já possuem uma metodologia de investigação bem definida e consolidada. Para o superfaturamento de obras públicas, por exemplo, a Polícia Federal possui um método de cálculo e classificação do superfaturamento proposto por Oliveira Lopes (2011). Entretanto, de acordo com Marcos Cavalcanti Lima (2021), para outros tipos de fraudes, como conluio e formação de cartel, a investigação é árdua e manual, além de depender muito da experiência do perito.

Pelas Leis 8.666 em Brasil (1993) e 14.133 em Brasil (2021), todo ato oficial relativo a uma licitação realizada por órgão ou entidade da Administração Pública Federal deve ser obrigatoriamente publicado no Diário Oficial da União (DOU). Por isso, o DOU é uma fonte rica de dados relativos às licitações que geralmente é utilizada durante as investigações de fraudes. Entretanto, a falta de estruturação e padronização das publicações, assim como a sua grande quantidade, em geral dezenas por licitação, dificultam o trabalho das pessoas envolvidas em uma investigação. Por esses e outros aspectos, estudos sobre a aplicação de

técnicas de aprendizado de máquina em publicações do DOU são escassos.

Em 2020, Marcos Lima et al. (2020) propuseram um modelo de classificação de publicações do DOU para detecção de fraudes e conluios em licitações de obras públicas no Brasil, como parte da ferramenta *Deep Vacuity*. Como descrito por Marcos Cavalcanti Lima (2021), essa ferramenta computacional tem como objetivo fornecer aos investigadores da Polícia Federal metodologias para a identificação de fraudes em licitações públicas através de técnicas de aprendizado de máquina.

Para que os modelos de aprendizado de máquina obtenham êxito na análise de textos e detecção de fraudes, são necessárias antes algumas tarefas a fim de tornar a análise mais eficiente e precisa, conhecidas como subtarefas. Uma das subtarefas mais importantes é o Reconhecimento de Entidades Nomeadas (REN), que tem como objetivo identificar e categorizar entidades textuais em um texto, como pessoas, organizações e localidades, por exemplo.

No contexto da ferramenta *Deep Vacuity*, o Reconhecimento de Entidades Nomeadas é de suma importância para a filtragem e vinculação de publicações. Atualmente, o reconhecimento de entidades nos textos das publicações do DOU é realizado por meio de correspondência direta, através do uso de expressões regulares. Todavia, como citado anteriormente, as publicações do DOU não possuem um padrão de escrita bem estruturado, o que faz com que a utilização de expressões regulares não seja o método mais eficiente. Em adição a não padronização das publicações, há também as complexidades inerentes ao processamento e interpretação da linguagem natural, como observado por Albanaz (2020). A presença de uma mesma palavra com significados diferentes (polissemia), de palavras diferentes com o mesmo significado (sinonímia) e de palavras pouco utilizadas (raridade) são complexidades da linguagem natural que exigem uma análise mais poderosa.

No campo do Processamento de Linguagem Natural (PLN), o *Transformer* desponta como a arquitetura mais poderosa atualmente, oferecendo uma capacidade poderosa de análise e geração de linguagem natural. Como observado por Wolf et al. (2020), nos últimos anos, o *Transformer* experimentou um crescimento exponencial, superando arquiteturas anteriormente predominantes no campo de PLN, como Redes Neurais Recorrentes (RNNs, do inglês *Recurrent Neural Networks*) e Convolucionais (CNNs, do inglês *Convolutional Neural Networks*) como apresentado em Salas, Barros Vidal e Martinez-Trinidad (2019). Esse crescimento é atribuído em grande parte ao sucesso de modelos de *Transformer* que, como observado por Prakash (2023), atingiram um desempenho jamais visto antes, como o BERT e o GPT, este último sendo o modelo por trás do ChatGPT.

As principais vantagens do *Transformer* em comparação com outras arquiteturas são a velocidade e eficiência de treinamento, possibilitadas pela facilidade na paralelização do treinamento, e a capacidade de capturar contextos de longo alcance, como descrito por Wolf et al. (2020). Além disso, o *Transformer* possibilita o pré-treinamento dos modelos em

uma quantidade massiva de textos genéricos, seguido pelo treinamento para tarefas mais específicas.

1.2 Objetivos

O objetivo primário desse trabalho é treinar modelos de *Transformer* pré-treinados para realizar o Reconhecimento de Entidades Nomeadas em publicações do Diário Oficial da União. Além disso, busca-se comparar o desempenho de diferentes modelos e investigar a relação entre as entidades e o desempenho dos modelos. De forma a delimitar o escopo desse trabalho, serão consideradas apenas publicações relativas a convênios.

Para permitir o atingimento do objetivo primário, um objetivo secundário a ser alcançado é a construção de uma base de publicações anotadas que conterá exemplos para o aprendizado dos modelos. Para isso, será realizado o mapeamento das entidades, a busca e validação automatizada das publicações, bem como a anotação das mesmas.

1.3 Organização do Trabalho

Este trabalho foi estruturado em seis capítulos. Nesse primeiro capítulo, são apresentadas as motivações e justificativas do trabalho, assim como os objetivos e organização do trabalho. O Capítulo 2 apresenta a fundamentação teórica dos principais tópicos abordados neste trabalho, como o Reconhecimento de Entidades Nomeadas e o *Transformer*. De forma a situar este trabalho dentro do contexto científico, no Capítulo 3 são revisados os trabalhos existentes sobre o tema em questão.

No Capítulo 4, é descrita a metodologia utilizada para a realização deste trabalho, juntamente com os resultados esperados. Os resultados obtidos e as discussões correspondentes serão apresentados no Capítulo 5. Por fim, no Capítulo 6, serão apresentadas as conclusões do estudo, juntamente com sugestões para pesquisas futuras.

2 Fundamentos Teóricos

O Processamento de Linguagem Natural (PLN) é uma subárea da Inteligência Artificial que reúne técnicas da Ciência da Computação e da Linguística para estudar a interação entre computadores e linguagens humanas, conforme descrito em Pustejovsky e Stubbs (2012). Para isso, o PLN se utiliza de técnicas de aprendizado de máquina e aprendizado profundo, e também de métodos estatísticos, para analisar, entender e gerar dados de linguagem humana em diversas formas, como texto, fala e imagens.

Como descrito em IBM (2023), escrever um *software* que entenda completamente o significado pretendido em dados de texto ou voz é extremamente complexo e difícil, dado a natureza ambígua da linguagem humana e também as suas diversas irregularidades, tais como homônimos, homófonos, sarcasmo, expressões idiomáticas, metáforas, exceções gramaticais e de uso, variações na estrutura da frase e entre outros. Por isso, para viabilizar o entendimento da linguagem humana pelo computador, o PLN gerou diversas subáreas, em que cada uma tem seu papel no entendimento da linguagem como um todo. A seguir estão listadas e descritas sucintamente algumas das principais subáreas do PLN:

- **Reconhecimento de fala** é uma subárea que tem como foco converter dados de voz em dados de texto. É utilizada em aplicações que precisam reconhecer comandos e perguntas no formato de voz, conforme definido em IBM (2023). Como descrito por Karatas (2023) e ilustrado pela Figura 2.1, um sistema de reconhecimento de fala é divido nas seguintes etapas: processamento do sinal de voz, onde ele é filtrado e convertido no melhor formato para análise; extração de características do sinal, como espectro de frequência, intensidade, tom, etc.; modelagem acústica, em que o objetivo é treinar um modelo de forma com que ele reconheça os sons e fonemas, e classifique-os de acordo com um alfabeto específico; modelagem de linguagem, que visa treinar um modelo para que ele compreenda a estrutura, organização e contexto da linguagem e torne o reconhecimento de fala mais coerente; e decodificação, que busca transformar o sinal de voz em texto. De acordo com IBM (2023), alguns dos maiores desafios dessa subárea são os diferentes sotaques, diferentes entonações, velocidade de fala, entre outros.

- **Análise de sintaxe e gramática** visa compreender a estrutura gramatical das sentenças e frases de um texto através da classificação gramatical das palavras, como define IBM (2023). O processo de classificação gramatical geralmente é uma subetapa de algumas tarefas em PLN, como análise sintática, rotulagem de função semântica e resumo de texto. Conforme Jurafsky e Martin (2023), pode-se dividir esta tarefa nas seguintes etapas: tokenização, que é o processo de dividir uma frase ou texto em

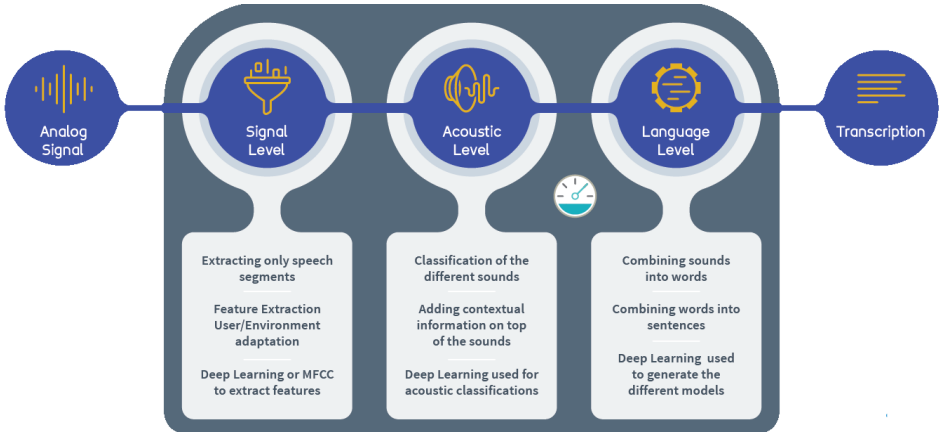


Figura 2.1 – Sistema de reconhecimento de fala.

Fonte: Alvarez (2017)

elementos individuais, chamados de *tokens*; resolução de ambiguidade, que busca resolver o problema em que uma palavra pode ter várias classificações gramaticais; e classificação, em que o objetivo é atribuir uma classe gramatical para cada *token*. Um exemplo de análise de sintaxe e gramática pode ser visto na Figura 2.2.

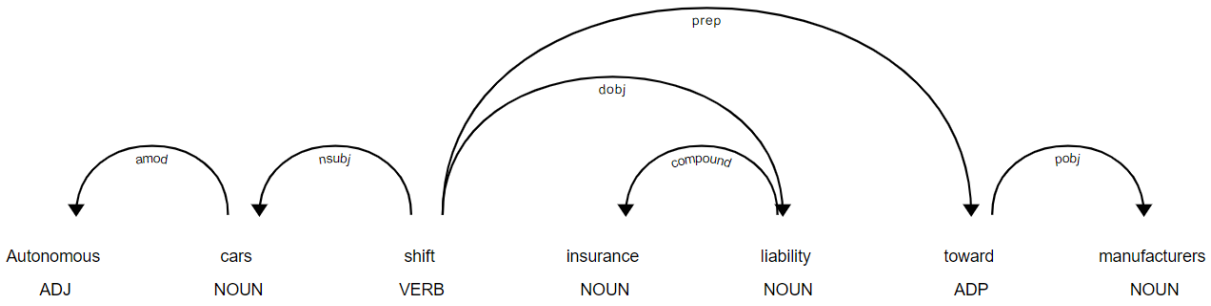


Figura 2.2 – Exemplo de análise de sintaxe e gramática.

Fonte: spaCy (2023b)

- **Análise semântica e pragmática** é a subárea que busca entender o significado das sentenças e frases em um dado contexto, incluindo a identificação de entidades, relacionamentos, intenções e desambiguação de sentido das palavras, como descrito em IBM (2023) e exemplificado na Figura 2.3. Esta subárea geralmente é um subetapa para tarefas como recuperação de informações, tradução automática e resumo de texto. O grande desafio da análise semântica e pragmática é a grande quantidade de palavras que possuem mais de um significado, por isso o contexto é muito importante nesta subárea.

- **Análise de sentimentos**, como define IBM (2023), tem como foco reconhecer sentimentos, emoções, sarcasmo, confusão e outras informações e qualidades subjetivas expressadas em um texto. Esta subárea é utilizada em aplicações como atendimento

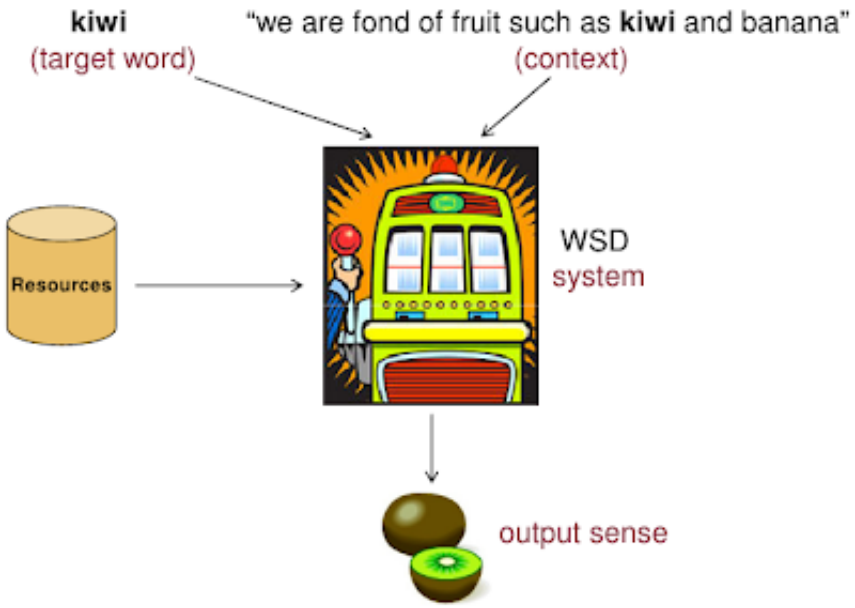


Figura 2.3 – Exemplo de desambiguação do sentido de uma palavra.

Fonte: Navigli (2023)

ao cliente, *marketing* e política a fim de entender e avaliar a opinião pública a respeito de serviços, produtos, pessoas, assuntos etc. Um exemplo de análise de sentimentos aplicada a uma frase pode ser visto na Figura 2.4. Geralmente, a análise de sentimentos envolve uma primeira etapa de pré-processamento do texto, uma segunda etapa de classificação do sentimento como sendo negativo, positivo ou neutro, e uma terceira etapa que busca refinar a classificação anterior, como descrito em Jurafsky e Martin (2023).

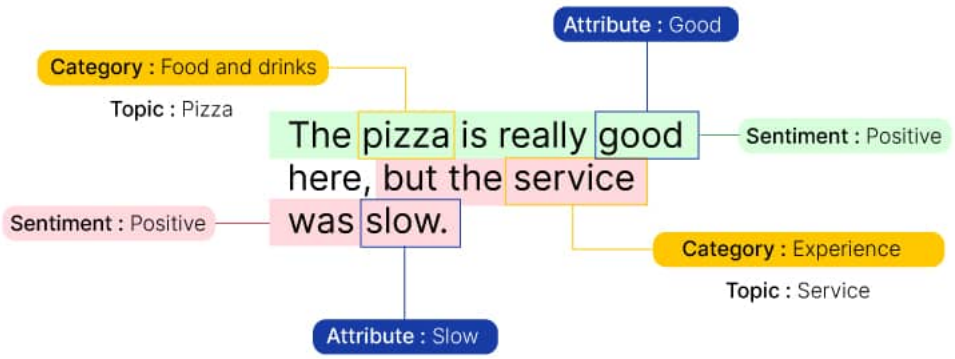


Figura 2.4 – Exemplo de análise de sentimento.

Fonte: Voxco (2023)

- **Geração de linguagem natural ou geração de texto** é a subárea que tem como objetivo realizar o oposto do reconhecimento de fala, ou seja, converter informação estruturada em linguagem natural, como definido em IBM (2023). O grande desafio desta subárea é gerar textos coerentes e gramaticalmente corretos a partir de dados de entrada. A geração de linguagem natural, geralmente, é uma subetapa de aplicações como *chatbots*, tradução de linguagem e resumo de texto. Um exemplo de um sistema de *chatbot* pode ser visualizado na Figura 2.5.

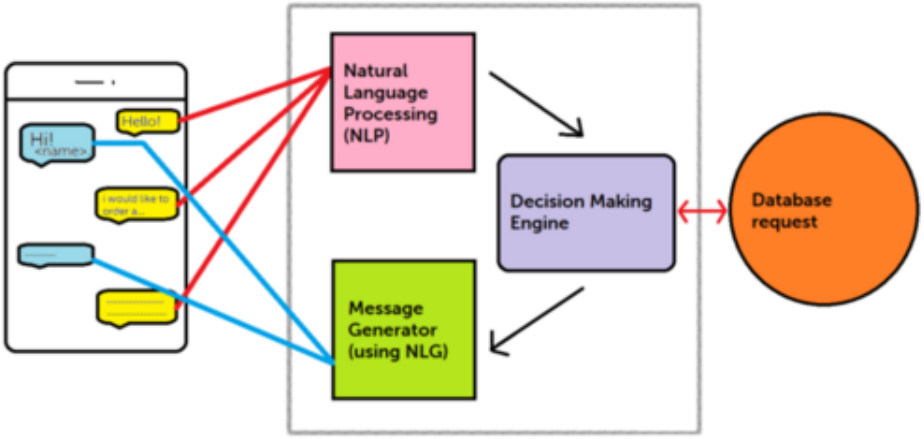


Figura 2.5 – Funcionamento de um chatbot.

Fonte: Ramesh (2019)

- **Reconhecimento de entidades nomeadas**, de acordo com Jurafsky e Martin (2023), se preocupa em identificar e classificar palavras ou frases como entidades nomeadas, tais como pessoas, organizações, localizações, datas, entre outras. O reconhecimento de entidades nomeadas é uma tarefa de extrema importância para outras subáreas como recuperação de informações, resposta a perguntas e resumo de texto. A Figura 2.6 mostra possíveis entidades de um texto.

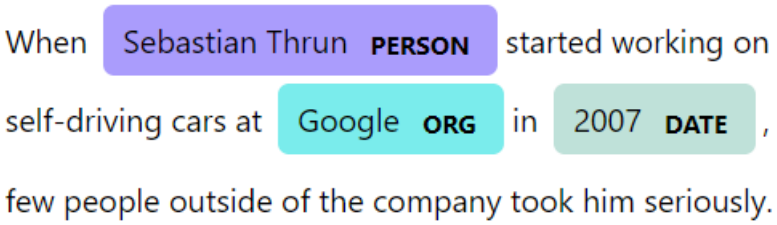


Figura 2.6 – Exemplo de um texto com entidades nomeadas.

Fonte: spaCy (2023b)

Este trabalho tem como foco a subárea de reconhecimento de entidades nomeadas e, por isso, detalha-se mais sobre ela na próxima seção. Com o intuito de promover a padroni-

zação e garantir a consistência das informações apresentadas, optou-se por embasar as três próximas seções em Jurafsky e Martin (2023).

2.1 Reconhecimento de Entidades Nomeadas (REN)

Como definido em Jurafsky e Martin (2023), uma entidade nomeada é tudo aquilo que pode ser referenciado como um substantivo próprio, como o nome de uma pessoa, uma localização ou uma organização. A tarefa de identificar as entidades nomeadas presentes em um texto é chamada de Reconhecimento de Entidades Nomeadas. Apesar da definição de entidade nomeada descrita anteriormente, segundo Jurafsky e Martin (2023) esta definição pode ser estendida de forma a incluir palavras que não são entidades propriamente ditas, como datas, horários, preços e entre outras.

Conforme ilustra a Tabela 2.1, extraída de Jurafsky e Martin (2023), existem algumas entidades genéricas que são típicas nas aplicações de REN. Entretanto, para alguns casos, é comum a utilização de entidades específicas para a aplicação em questão. Um exemplo disto, e que será mostrado mais adiante, são as entidades utilizadas neste trabalho, que são específicas de convênios públicos.

Tabela 2.1 – Entidades nomeadas genéricas.

Tipo	**Rótulo**	**Exemplos de categorias**	**Exemplo de frase**
Pessoas	PER	pessoas, personagens	**Turing** é um gigante da ciência da computação.
Organização	ORG	empresas, times de esportes	O **IPCC** alertou sobre o ciclone.
Localização	LOC	regiões, montanhas, mares	**Mt. Sanitas** fica em **Sunshine Canyon**.
Entidade geopolítica	GPE	países, estados	**Palo Alto** está aumentando as taxas de estacionamento.

Fonte: Jurafsky e Martin (2023)

Em geral, a tarefa de reconhecer entidades nomeadas é uma subetapa muito importante para outras tarefas em PLN, como análise de sentimentos, resposta à perguntas e análise semântica. O grande desafio do REN é a ambiguidade no sentido das palavras, visto que uma mesma palavra pode ter vários significados e sua classificação é dependente do contexto. Um exemplo de ambiguidade pode ser visto na Figura 2.7, onde a palavra "Vasco da Gama" é rotulada como três tipos diferentes de entidade.

Resumidamente, o REN geralmente envolve as etapas de tokenização, reconhecimento das entidades nomeadas e, dependendo da técnica utilizada, a vinculação das entidades nomeadas. A definição de cada uma das etapas pode ser vista a seguir.

1. **Tokenização:** esta etapa envolve segmentar o texto em palavras individuais ou *tokens*, que são as unidades de processamento em PLN;

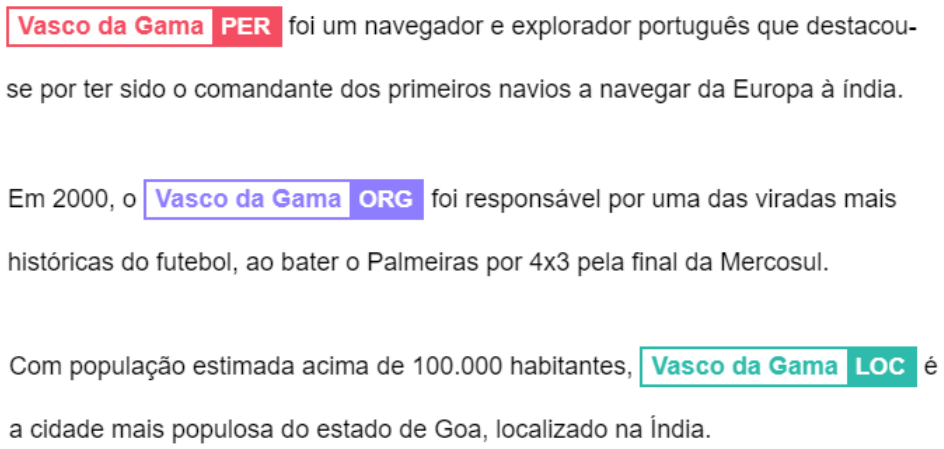


Figura 2.7 – Exemplo de trechos com entidades nomeadas ambíguas.

2. **Reconhecimento das entidades nomeadas:** envolve a identificação e a classificação das entidades nomeadas no texto;
3. **Vinculação das entidades nomeadas:** busca vincular as entidades nomeadas a uma base de conhecimento ou ontologia para desambiguar e extrair informações adicionais sobre as entidades.

O reconhecimento de entidades nomeadas em um texto pode ser feito por meio de métodos baseados em regras e métodos baseados em estatística (*machine learning*).

2.1.1 Métodos Baseados em Regras

Os métodos baseados em regras são os métodos mais simples de reconhecimento de entidades. Nos métodos baseados em regras, o REN utiliza um conjunto de regras pré-definidas para identificar e classificar as entidades. Em geral, esses métodos são utilizados para o reconhecimento de entidades simples e que, geralmente, possuem um padrão de escrita. Para entidades mais complexas, que possuem inúmeras formas de escrita e dependem do contexto, os métodos baseados em regras são limitados. A seguir são apresentados alguns dos principais métodos baseados em regras.

2.1.1.1 Expressões Regulares

O método baseado em regras mais famoso e comumente utilizado é a expressão regular. De acordo com Jurafsky e Martin (2023), uma expressão regular é uma notação algébrica que busca caracterizar um conjunto de strings. Essa expressão regular é então utilizada para buscar este conjunto de strings em textos.

Uma expressão regular pode ser um simples caractere ou uma sequência de caracteres. Por exemplo, para procurar o caractere "V"em um texto, basta utilizar a expressão */V/*. Para procurar a palavra "banana", utiliza-se a expressão */banana/*. As expressões regulares são sensíveis às letras maiúsculas e minúsculas, portanto */banana/* é diferente de */Banana/*. Para solucionar isto, basta a utilização de chaves, que especifica uma disjunção de caracteres. Logo, a expressão */[Bb]anana/* retornaria tanto "Banana" quanto "banana". Para retornar qualquer digito, por exemplo, utiliza-se a expressão */[1234567890]/*. Entretanto, para casos em que há uma sequência bem definida, utiliza-se um traço entre o primeiro e o último elemento da sequência para indicar todos os caracteres da sequência. Para o caso anterior, por exemplo, a expressão */0-9/* também retornaria qualquer digito.

A utilização do acento circunflexo logo após a chave de abertura indica negação. Por exemplo, a expressão */ˆA-Z/* significa que qualquer caractere pode ser retornado, exceto qualquer letra maiúscula. Caracteres opcionais também podem ser definidos por meio do ponto de interrogação "?". Por exemplo, a expressão */maçãs?/* significa que pode ser retornado tanto "maçã"quanto "maçãs, pois o "?" indica que o caractere precedente é opcional. Para algumas expressões mais comuns, existem operadores específicos, como pode ser visto na Tabela 2.2.

Tabela 2.2 – Operadores específicos para expressões regulares comuns.

Regex	**Expansão**	**Correspondência**	**Exemplos**
\d	[0-9]	qualquer digito	Festa de 5 anos.
\D	[ˆ0-9]	qualquer não dígito	Lua azul.
\w	[a-zA-z0-9_]	qualquer alfanumérico/sublinhado	Daiyu.
\W	[ˆ\w]	um não alfanumérico	!!!!
\s	[\r\t\n\f]	espaço em branco (espaço, tabulação)	Em Concórdia.
\S	[ˆ\s]	sem espaço em branco	Em Concórdia.

Fonte: Jurafsky e Martin (2023)

O uso de expressões regulares também permite especificar o número de ocorrências de uma expressão. Existem caracteres que especificam os números de ocorrências mais comuns, entretanto o número de ocorrências pode ser especificado por */{n,m}/*, em que o *n* especifica a quantidade mínima de ocorrências e o *m* especifica a quantidade máxima. As variações dessa expressão regular, assim como os caracteres especiais que especificam números de ocorrências mais comuns, podem ser vistos na Tabela 2.3.

Por serem especiais ou terem alguma função dentro das expressões regulares, alguns caracteres devem ser precedidos de barra invertida quando quando estes fazem parte da expressão em que deseja-se encontrar. A Tabela 2.4 mostra alguns desses caracteres.

2.1.1.2 Correspondência Baseada em Dicionário

A correspondência baseada em dicionário é um método baseado em regras que consiste na utilização de um dicionário ou lista pré-definida de entidades nomeadas a fim de

Tabela 2.3 – Expressões regulares para especificar o número de ocorrências.

Regex	Correspondência
*	zero ou mais ocorrências do caractere ou expressão anterior
+	uma ou mais ocorrências do caractere ou expressão anterior
?	zero ou uma ocorrência do caractere ou expressão anterior
{n}	exatamente *n* ocorrências do caractere ou expressão anterior
{n,m}	de *n* a *m* ocorrências do caractere ou expressão anterior
{n,}	ao menos *n* ocorrências do caractere ou expressão anterior
{,m}	até *m* ocorrências do caractere ou expressão anterior

Fonte: Jurafsky e Martin (2023)

Tabela 2.4 – Caracteres que precisam ser precedidos de barra invertida.

Regex	Correspondência	Exemplos
*	um asterisco "*"	"K*A*P*L*A*N"
\.	um ponto "."	"Dr. Livingston, eu presumo"
\?	um ponto de interrogação	"Por que eles não vêm e dão uma mão?"
\n	uma nova linha	
\t	uma tabulação	

Fonte: Jurafsky e Martin (2023)

classificá-las em um texto. Esse dicionário de entidades nomeadas pode ser criado manualmente ou gerado automaticamente a partir de um conjunto de textos. Para cada entrada do dicionário, ou seja, para cada entidade nomeada, geralmente define-se uma lista de variações da mesma, como abreviações, acrônimos, diferentes ortografias, etc. Basicamente, o método de correspondência baseada em dicionário procura no texto todas palavras que correspondem às entradas no dicionário e, caso haja correspondência, classifica a palavra como entidade nomeada com base em suas entradas de dicionário correspondentes. Por exemplo, se um texto contém a palavra "Microsoft" e esta palavra está incluída no dicionário como nome de uma organização, então a palavra "Microsoft" seria classificada como uma entidade nomeada do tipo "organização".

Esse método é utilizado em casos onde as entidades nomeadas são bem definidas e conhecidas e também possui a vantagem de ser computacionalmente eficiente e fácil de utilizar, visto que não requer a utilização de modelos de *machine learning* e grandes *datasets* de treinamento. Em contrapartida, é um método limitado quanto à ambiguidade no sentido das palavras e a erros ortográficos. Por isso, este método é utilizado em conjunto com outros métodos de REN para melhorar a acurácia do sistema. Neste trabalho, a correspondência baseada em dicionário será utilizada em conjunto com modelo de *deep learning* para a tarefa de REN.

2.1.1.3 Regras Gramaticais

Para a identificação de entidades nomeadas, o método baseado em regras gramaticais utiliza a gramática livre de contexto (CFG, do inglês *context-free grammar*), que consiste

em um conjunto de regras gramaticais pré-definidas que definem a estrutura sintática e semântica de uma linguagem. Um exemplo de regra gramatical pode ser visto na Figura 2.8.

Ao identificar entidades nomeadas em um texto por meio da sintaxe, esse método analisa a estrutura sintática do texto e identifica as entidades nomeadas de acordo com o conjunto de regras gramaticais pré-definidas. Essas regras podem especificar aspectos como classe gramatical, capitalização, precedente e subsequente. As regras gramaticais também podem ser utilizadas para identificar entidades nomeadas por meio da semântica do texto. Por exemplo, entidades podem ser identificadas baseadas em suas preposições.

Uma vantagem do método baseado em regras gramaticais é que nele é possível considerar a estrutura e o contexto das entidades nomeadas sistematicamente. Entretanto, assim como todo método baseado em regras, o método baseado em regras gramaticais é dependente da acurácia das regras e estas, na maioria das vezes, precisam ser feitas manualmente para garantir o bom desempenho do método.

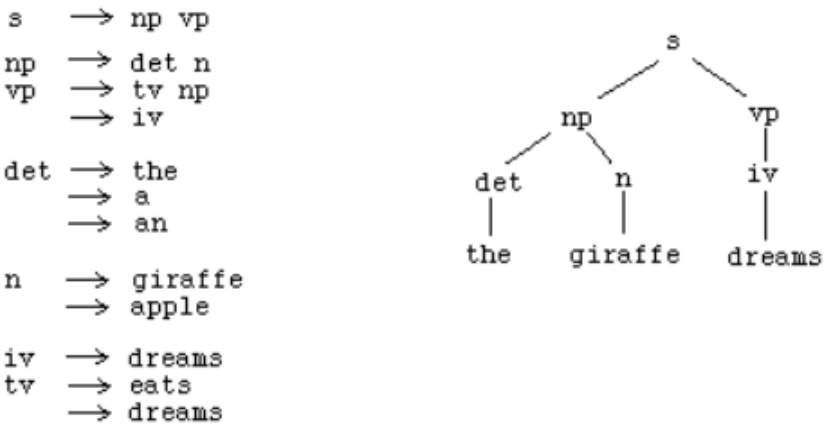


Figura 2.8 – Uma gramática e uma árvore de análise para "*the giraffe dreams*".

Fonte: Tucker (2002)

2.1.2 Métodos Baseados em Estatística (*Machine Learning*)

Os métodos baseados em estatística se baseiam em técnicas de estatística aplicadas a um grande conjunto de textos. Estes métodos se apoiam na teoria de que as entidades nomeadas possuem um certo padrão e regularidade nos textos e que, ao modelar esse padrão estatisticamente, é possível identificar e classificar as entidades nomeadas com alto grau de acurácia. Apesar de serem uma evolução em relação aos métodos baseados em regras, estes métodos ainda possuem limitações quanto às entidades que dependem do contexto ou que não são abrangidas no conjunto de textos utilizados para a aplicação das técnicas estatísticas. A seguir são apresentados alguns dos principais métodos baseados em estatística.

2.1.2.1 Entropia Máxima

O REN baseado na entropia máxima atribui um valor para cada palavra ou *token* do texto que indica a probabilidade de ser uma entidade nomeada. A atribuição deste valor é realizada pelo modelo de entropia máxima, que é previamente treinado com um conjunto de textos anotados no qual as entidades nomeadas são rotuladas de acordo com seu tipo.

Neste método, a probabilidade é estimada por um algoritmo iterativo chamado de princípio da entropia máxima. De acordo com Jurafsky e Martin (2023), a entropia é uma medida de informação, conforme Equação 2.1:

$$H(X) = -\sum_{x \in \chi} p(x) \log_2 p(x). \tag{2.1}$$

Onde uma dada variável aleatória X, que varia sobre o que deseja-se prever (conjunto χ) e com uma função de probabilidade $p(x)$.

Diferentemente dos métodos baseados em regras, o método da entropia máxima consegue lidar com características linguísticas mais avançadas, como o contexto, e por isso é mais acurado e eficiente. Além disso, por ser um método estatístico, possui como vantagem a escalabilidade, visto que grandes conjuntos de textos podem ser utilizados para treinar o modelo e torná-lo mais eficiente. A flexibilidade também é uma característica importante deste método, visto que é possível estendê-lo a várias linguagens.

Uma limitação dos métodos baseados em estatísticas é que, em geral, requerem uma grande quantidade de textos anotados para a produção de modelos eficientes. Além disso, se comparados com os métodos baseados em regras, o poder computacional e o tempo de execução requeridos são maiores.

2.1.2.2 Modelo Oculto de Markov

Conforme definido em Jurafsky e Martin (2023), um Modelo Oculto de Markov (HMM, do inglês *Hidden Markov Model*) é um clássico modelo de sequência probabilística, em que para cada unidade de um texto (letras, palavras, *tokens*, etc.), é calculada uma distribuição de probabilidade sobre todos os possíveis rótulos, e o que possui maior probabilidade é escolhido para rotular a unidade. De forma resumida, o HMM baseia-se nas dependências entre palavras adjacentes para prever o rótulo de cada palavra.

O HMM é baseado nas cadeias de Markov, que basicamente são representadas por grafos, onde os estados podem ser palavras ou símbolos e as transições representam as probabilidades, e a soma de todas elas deve ser igual a um. A Figura 2.9 mostra dois exemplos de cadeias de Markov.

Uma cadeia de Markov é composta por três componentes: um conjunto de N estados ($Q = q_1 q_2 ... q_N$); uma matriz de probabilidade de transição ($A = a_{11} a_{12} ... a_{N1} ... a_{NN}$), em

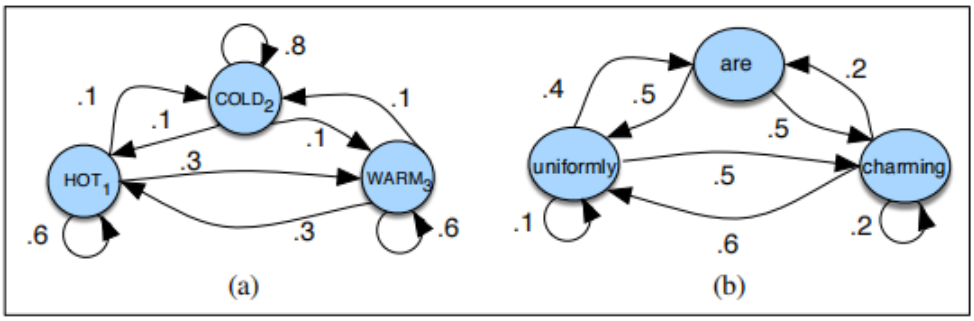


Figura 2.9 – Uma cadeia de Markov para clima (a) e outra para palavras (b), mostrando estados e transições.

Fonte: Jurafsky e Martin (2023)

que cada a_{ij} representa a probabilidade de transição do estado i para o estado j; e uma distribuição de probabilidade inicial ($\pi = \pi_1,\pi_2,...,\pi_N$). A cadeia é utilizada quando é calculada a probabilidade de eventos observáveis, ou seja, palavras ou símbolos que são observados diretamente no texto. Para eventos ocultos, em que rótulos devem ser inferidos a partir de palavras do texto, como entidades nomeadas por exemplo, utiliza-se o HMM. De acordo com Jurafsky e Martin (2023), o HMM é especificado pelos três componentes que especificam uma cadeia de Markov, adicionado de dois componentes: uma sequência de T observações ($O = o_1 o_2 ... o_T$), cada uma extraída de um vocabulário $V = v_1,v_2,...,v_V$; e uma sequência de probabilidades de observação ($B = b_i(o_t)$), também chamadas de probabilidades de emissão, cada uma expressando a probabilidade de uma observação ser gerada a partir de um estado q_i

O HMM possui as mesmas vantagens que o método de entropia máxima, com o benefício de obter uma eficiência igual ou maior com *datasets* relativamente menores. Uma das limitações deste método é que ele assume que todos rótulos são independentes de um do outro, o que nem sempre é verdade.

2.1.2.3 Campos Aleatórios Condicionais

Um Campo Aleatório Condicional (CRF, do inglês *Conditional Random Field*) é um modelo de sequência discriminativa baseado em modelos log-lineares, como define Jurafsky e Martin (2023). Formalmente, seja $x_{1:N}$ as observações (palavras em um texto), e $z_{1:N}$ os rótulos ocultos (entidades nomeadas). De acordo com Zhu (2007), um CRF define a seguinte probabilidade condicional:

$$p(z_{1:N}|x_{1:N}) = \frac{1}{Z} exp\left(\sum_{n=1}^{N}\sum_{i=1}^{F}\lambda_i f_i(z_{n-1}, z_n, x_{1:N}, n)\right) \tag{2.2}$$

em que o escalar Z é o fator de normalização, e é definido como a soma do número exponencial de sequências:

$$Z = \sum_{z_{1:N}} exp\left(\sum_{n=1}^{N}\sum_{i=1}^{F}\lambda_i f_i(z_{n-1}, z_n, x_{1:N}, n)\right) \tag{2.3}$$

Um importante componente do CRF são as funções de recurso. Uma função de recurso mapeia um recurso de entrada particular (palavra de um texto) e seu rótulo correspondente (entidade nomeada) em um valor de recurso real ou binário com o objetivo de capturar a informação contextual de cada palavra e suas palavra vizinhas. Basicamente, a função relaciona os recursos de entrada e os rótulos de saída em um modelo probabilístico. Como um exemplo, Zhu (2007) define uma simples função de recurso em que se a palavra atual é "John" e o estado atual é "PESSOA", ela produz um valor binário igual a 1:

$$f_1(z_{n-1}, z_n, x_{1:N}, n) = \begin{cases} 1 & \text{se } z_n = \text{PESSOA e } x_n = \text{John} \\ 0 & \text{caso contrário} \end{cases} \tag{2.4}$$

Uma função de recurso muito conhecida e utilizada principalmente para localidades é um *gazetteer*, que é uma lista que geralmente contém milhares de nomes de lugares com informações políticas e geográficas detalhadas.

Uma vantagem do CRF é a capacidade de lidar com recursos sobrepostos e complexos, tais como combinações de palavras e informações contextuais. Isso possibilita a consideração mais eficiente de propriedades sintáticas e semânticas das entidades nomeadas.

2.1.3 Métodos Baseados em *Deep Learning*

Apesar dos métodos baseados em regras e em estatística terem contribuído significativamente para a evolução da tarefa de REN, os métodos baseados em *deep learning* revolucionaram essa área. Esses métodos são capazes de identificar padrões complexos nos textos e alcançar um desempenho de estado da arte. Abaixo estão listadas as principais técnicas de *deep learning* utilizadas para a tarefa de REN:

- **Redes Neurais Recorrentes:** como descrito em Jurafsky e Martin (2023), as RNNs (do inglês *Recurrent Neural Networks*) são redes neurais que contém um ciclo dentro de suas conexões de rede, ou seja, o valor de alguma unidade é direta ou indiretamente dependente de suas próprias saídas anteriores como uma entrada. Esse *feedback* cíclico permite com que dependências entre palavras sejam identificadas. As RNNs mais populares são a LSTM (do inglês *Long Short-Term Memory*) e a GRU (do inglês *Gated Recurrent Unit*);
- **Redes Neurais Convolucionais:** as CNNs (do inglês *Convolutional Neural Networks*), diferentemente das RNNs, são redes neurais do tipo *feedfoward*. Elas consistem de

múltiplas camadas de nós interconectados que detectam padrões nos dados de entrada. Em geral, essas redes são muito utilizadas para tarefas de reconhecimento de imagem e visão computacional, mas também podem ser aplicadas a textos;

- **Modelos Baseados em *Transformers*:** os modelos que mais obtiveram sucesso e alcançaram o estado da arte na tarefa de REN foram os baseados em *Transformers*. Esses modelos são baseados em um mecanismo de autoatenção que permite com o contexto global dos textos de entrada seja considerado na tarefa de REN. Dois desses modelos serão utilizados nesse trabalho e serão detalhados mais a frente;
- **Modelos Híbridos:** os Modelos Híbridos combinam multiplos modelos de *deep learning* ou combinam técnicas de *deep learning* com outros tipos de técnicas, como as de *machine learning*, de forma a aproveitar os pontos fortes de cada técnica. Um Modelo Híbrido poderia, por exemplo, utilizar RNN para identificar o contexto local e um modelo baseado em *Transformer* para identificar o contexto global.

Modelos baseados em *deep learning* necessitam que uma grande quantidade de textos anotados sejam utilizados para o treinamento de forma a alcançar resultados satisfatórios. Além disso, são modelos que exigem um poder computacional considerável, onde alguns necessitam de unidades de processamento gráfico (GPUs, do inglês *Graphics Processing Units*) de grande capacidade, como os modelos baseados em *Transformers*, que serão detalhados a seguir.

2.2 *Transformers*

O *Transformer* é um tipo de arquitetura de rede neural que foi projetada inicialmente em Vaswani et al. (2017) para tarefas de tradução de texto. Como descrito em Jurafsky e Martin (2023), o *Transformer* traz como novidade dois mecanismos principais: a autoatenção e as codificações posicionais. Estes mecanismos permitem relacionar as palavras separadas por longas distâncias de forma mais eficiente e, consequentemente, melhoram de forma significante o entendimento do contexto global de um texto. Antes do desenvolvimento da arquitetura de *Transformer*, as RNNs eram o tipo de arquitetura mais comum aplicado às tarefas de PLN e também podiam lidar com informações distantes. Entretanto, o grande diferencial do *Transformer* é a paralelização, que permitiu com que modelos fossem treinados com uma quantidade muito maior de dados e mais rapidamente.

Basicamente, os *Transformers* mapeiam sequências de vetores de entrada $(x_1,...,x_n)$ em sequências de vetores de saída $(y_1,...,y_n)$ do mesmo tamanho, como definido em Jurafsky e Martin (2023). Os *Transformers* são compostos por pilhas de blocos de *transformer*, em que cada bloco consiste na combinação de camadas de normalização, redes *feedforward*,

camadas de autoatenção e conexões residuais, como pode ser visto na Figura 2.10. A seguir, cada um destes componentes detalhados.

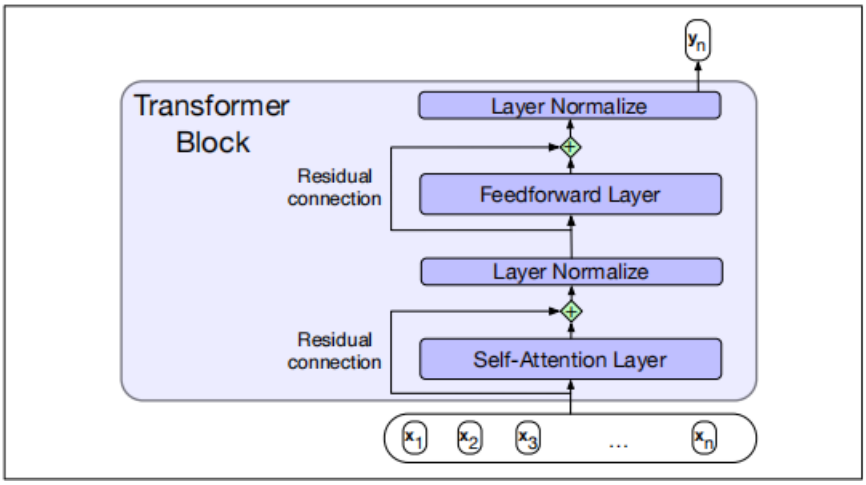


Figura 2.10 – Um bloco de *transformer* e suas camadas.

Fonte: Jurafsky e Martin (2023)

2.2.1 Redes *feedforward*

Apesar de não terem sido introduzidas com o *Transformer*, as redes *feedforward* possuem um papel importante para a arquitetura. De acordo com Jurafsky e Martin (2023), se tratam das redes neurais mais simples, em que os nós são conectados sem ciclos, ou seja, a saída de cada nó em cada camada é passada apenas para os nós da próxima camada. Como pode ser visto na Figura 2.11, uma rede *feedforward* simples é composta de três tipos de nós: nós de entrada, nós ocultos e nós de saída. Quando cada nó de uma camada recebe como entrada a saída de todos os nós da camada anterior, a camada é dita totalmente conectada. Os nós ocultos h_i fazem parte do núcleo da rede neural e realizam uma somada ponderada em suas entradas seguida da aplicação de uma não linearidade. A saída da camada oculta, o vetor h, é dada por:

$$h = \sigma\left(Wx + b\right). \tag{2.5}$$

Onde σ é uma função de ativação, que neste exemplo é a função sigmoide, W é a matriz de pesos da camada oculta, x é o vetor de entrada e b é o vetor de *bias*. O vetor h é então tomado como entrada para determinar a saída intermediária z:

$$z = Uh. \tag{2.6}$$

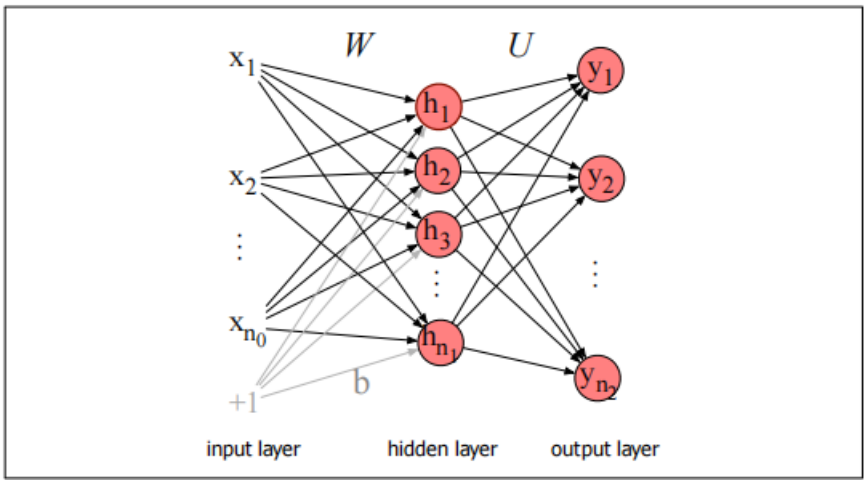


Figura 2.11 – Uma simples rede neural *feedforward*.

Fonte: Jurafsky e Martin (2023)

Onde U é a matriz de pesos da camada de saída. De forma a converter o vetor z, que é um vetor de números reais, em um vetor de probabilidades, é usualmente aplicada a função *softmax*, como indicado na Equação 2.7:

$$y = softmax(z) = \frac{exp(z_i)}{\sum_{j=1}^{d} exp(z_j)} \; 1 \leq i \leq d. \tag{2.7}$$

Onde d é a dimensão do vetor z. Portanto, uma rede neural *feedforward* é definida pelas Equações 2.5, 2.6 e 2.7.

2.2.2 Autoatenção

De acordo com Jurafsky e Martin (2023), uma abordagem baseada no mecanismo de atenção é composta por um conjunto de comparações com itens relevantes em algum contexto, uma normalização dessas pontuações para fornecer uma distribuição de probabilidade, seguida por uma soma ponderada usando essa distribuição. No exemplo da Figura 2.12, o cálculo da saída y_3 é baseada em um conjunto de comparações entre a entrada atual x_3 e as entradas precedentes, ou seja, x_1 e x_2. Como descrito em Jurafsky e Martin (2023), a forma mais simples de comparação entre duas entradas, em uma camada de autoatenção, é o produto escalar, que gera uma pontuação:

$$score(x_i, x_j) = x_i \cdot x_j. \tag{2.8}$$

No exemplo em questão, o conjunto de comparações para a saída y_3 geraria três pontuações: $x_3 \cdot x_1$, $x_3 \cdot x_2$ e $x_3 \cdot x_3$. Quanto maior a pontuação, mais similares os vetores são.

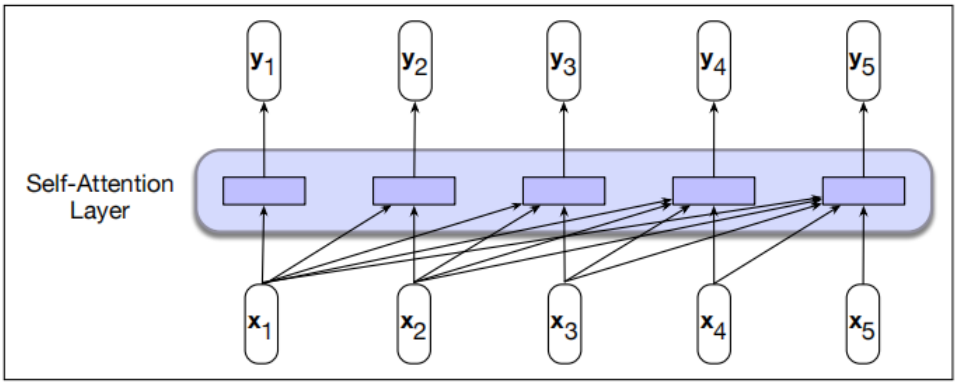


Figura 2.12 – Fluxo de informação em um modelo de autoatenção simples.

Fonte: Jurafsky e Martin (2023)

Após o conjunto de comparações, essas pontuações são normalizadas a fim de se obter uma distribuição de probabilidade. Essa normalização com uma função *softmax*, que retorna um vetor de pesos α_{ij} que indica a relevância de cada uma das entradas precedentes x_j na entrada atual x_i, como indicada na Equação 2.9:

$$\alpha_{ij} = softmax(score(x_i, x_j)) = \frac{exp(score(x_i, x_j))}{\sum_{k=1}^{i} exp(score(x_i, x_k))} \forall j \leq i. \tag{2.9}$$

Dada a normalização, a saída atual y_i é calculada por meio da soma ponderada das entradas precedentes até a entrada atual x_i, em que os pesos são dados pelo vetor de pesos α:

$$y_i = \sum_{j \leq i} \alpha_{ij} x_j. \tag{2.10}$$

O funcionamento da camada de autoatenção em um *Transformer* é parecido com o que foi mostrado anteriormente, entretanto é tratado de uma forma mais sofisticada. Neste caso, existem três funções as quais cada entrada exerce durante o processo de atenção, são elas:

- *Query* (consulta): É a entrada atual, ou seja, o foco atual do processo de atenção, o que está sendo comparado com as entradas precedentes;
- *Key* (chave): É a entrada precedente que está sendo comparada com a entrada atual;
- *Value* (valor): é o valor utilizado para computar a saída para a entrada atual.

Com isso, são introduzidas as matrizes de pesos W^Q, W^K e W^V, que são utilizadas para projetar cada vetor de entrada x_i em uma representação de sua função como chave, consulta ou valor:

$$q_i = W^Q x_i;\ \ k_i = W^K x_i;\ \ v_i = W^V x_i. \tag{2.11}$$

A partir dessas projeções, a comparação da entrada atual x_i com uma entrada precedente x_j passa a ser o produto escalar entre seu vetor consulta q_i e o vetor chave k_j do elemento precedente (veja a Equação 2.12).

$$score(x_i, x_j) = \frac{q_i \cdot k_j}{\sqrt{d_k}}. \tag{2.12}$$

Onde o d_k é a dimensão dos vetores consulta e chave, e a divisão da pontuação pelo fator $\sqrt{d_k}$ tem como objetivo evitar problemas numéricos e uma perda efetiva de gradientes durante o treinamento, como descrito em Jurafsky e Martin (2023). Assim como na Equação 2.9, essa pontuação também é passada pela função *softmax* e, em seguida, a saída atual é calculada com base em uma soma ponderada, como indicado na Equação 2.13.

$$y_i = \sum_{j \leq i} \alpha_{ij} v_j. \tag{2.13}$$

Em que os pesos são dados pelo vetor valor v. O mesmo exemplo da Figura 2.12 pode ser visto na Figura 2.13, porém com o modelo de autoatenção utilizado em um *Transformer*.

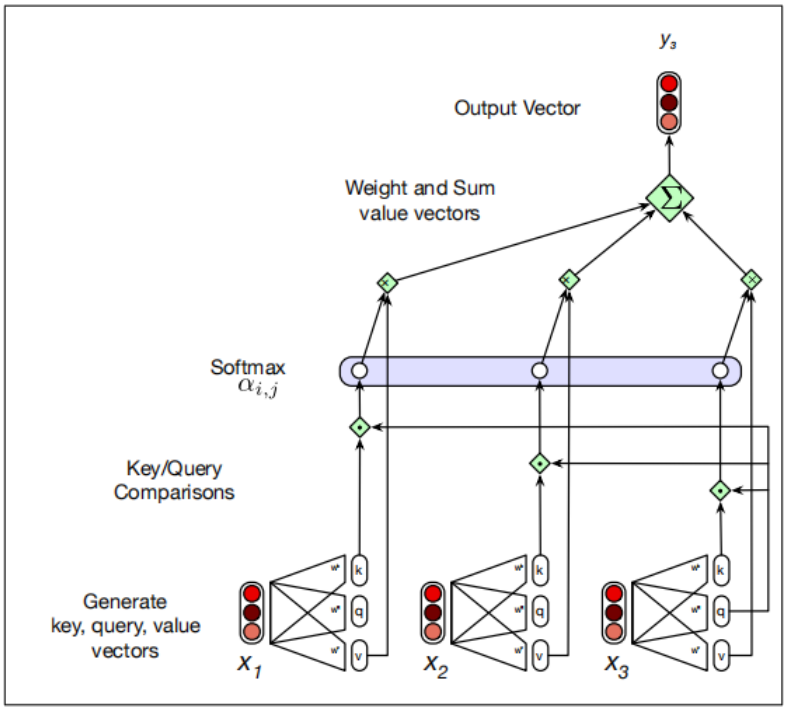


Figura 2.13 – Fluxo de informação em um modelo de autoatenção em um *Transformer*.

Fonte: Jurafsky e Martin (2023)

Todo o processo de autoatenção explicado anteriormente foi feito para o cálculo de um único valor de saída em um único passo de tempo. Entretanto, visto que o cálculo de cada saída é independente, é possível paralelizar este processo por meio da utilização de matrizes. Ao representar todas as entradas em uma matriz X, a Equação 2.11 toma a forma mostrada na Equação 2.14:

$$Q = XW^Q; \;\; K = XW^K; \;\; V = XW^V. \tag{2.14}$$

Ao transformar as Equações 2.12 e 2.13 em equações de matrizes e combiná-las, o processo de autoatenção pode ser reduzido a seguinte equação:

$$AutoAtencao(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{2.15}$$

Como o termo QK^T calcula não somente as comparações das entradas precedentes, mas também das entradas subsequentes, os elementos da porção triangular superior da matriz QK^T são computados como $-\infty$, o que elimina qualquer conhecimento das palavras seguintes na sequência, como descrito em Jurafsky e Martin (2023).

2.2.3 Conexões Residuais

Em redes neurais profundas, as conexões residuais transportam informação de uma camada inferior para uma camada superior sem passar pelas camadas intermediárias. De acordo com He et al. (2016), as conexões residuais permitem com que camadas superiores tenham acesso direto às informações das camadas inferiores, o que melhora o processo de aprendizado.

Em um *Transformer*, a conexão residual consiste em passar adiante para a próxima camada o vetor de saída da camada atual somado ao vetor de entrada da camada atual. Na Figura 2.10 são utilizadas conexões residuais nas subcamadas de autoatenção e *feedforward*.

2.2.4 Normalização

A camada de normalização, introduzida em Ba, Kiros e Hinton (2016), tem como objetivo melhorar o desempenho do treinamento ao manter os valores das camadas ocultas em uma faixa de valores favorável ao treinamento baseado em gradiente, conforme explicado em Jurafsky e Martin (2023). Dada uma camada oculta com dimensionalidade d_h, o processo de normalização se inicia com o cálculo da média e do desvio padrão dos elementos do vetor a ser normalizado, como se segue:

$$\mu = \frac{1}{d_h} \sum_{i=1}^{d_h} x_i, \tag{2.16}$$

$$\sigma = \sqrt{\frac{1}{d_h} \sum_{i=1}^{d_h} (x_i - \mu)^2}. \tag{2.17}$$

Onde μ é a média, σ é o desvio padrão e x_i é o i-ésimo elemento do vetor x a ser normalizado. Após isso, os elementos do vetor x são normalizados por meio da subtração pela média e em seguida da divisão pelo desvio padrão:

$$\hat{x} = \frac{x - \mu}{\sigma}. \tag{2.18}$$

Onde $\hat{x}$ é o vetor x normalizado com média igual a zero e desvio padrão igual a 1. Na implementação típica da camada de normalização, um par de parâmetros que podem ser aprendidos, γ e β, são introduzidos para representar os valores de ganho e deslocamento, respectivamente. Esses parâmetros desempenham um papel crucial no processo de normalização e se relacionam com o vetor normalizado da seguinte forma:

$$CamadaNorm = \gamma \hat{x} + \beta. \tag{2.19}$$

2.2.5 Codificação Posicional

Uma outra inovação trazida pelo *Transformer* é o conceito de codificação posicional. Dado que o *Transformer* não possui recorrência e nem convolução, ele não é capaz de por si só compreender a noção de posição relativa ou absoluta dos *tokens* de um texto de entrada.

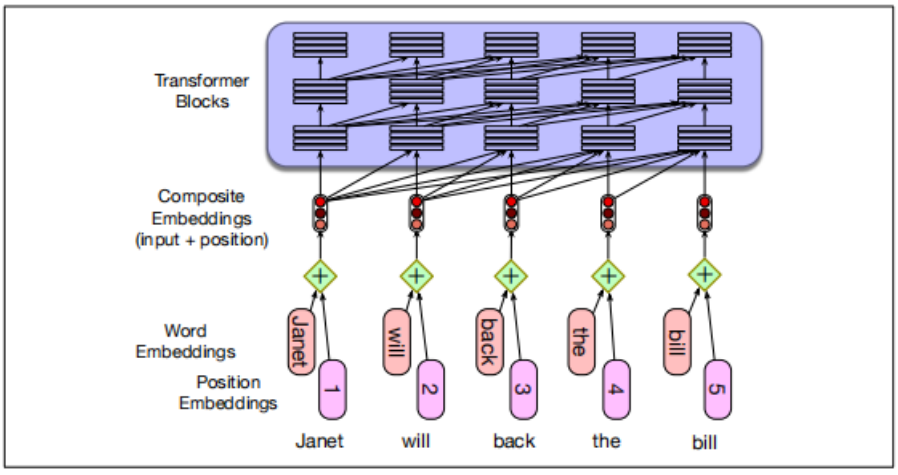


Figura 2.14 – Exemplo de codificação posicional.

Fonte: Jurafsky e Martin (2023)

Para resolver isso, essa informação é armazenada nos próprios *tokens* por meio da soma de codificações posicionais, como ilustra a Figura 2.14. Com isso, durante o treinamento, o modelo aprende como interpretar essas codificações posicionais.

Existem diversas escolhas de codificações posicionais, sejam elas aprendidas ou fixas. Em Vaswani et al. (2017), são utilizadas funções seno e cosseno de diferentes frequências, com o argumento de que essas funções permitiriam com que o modelo aprenderia mais facilmente as posições relativas. Apesar das funções seno e cosseno obterem um bom desempenho, o desenvolvimento de melhores representações de posição ainda é um tópico em constante estudo, como descrito em Jurafsky e Martin (2023).

2.3 Métricas de Avaliação

A fim de avaliar o desempenho e a qualidade dos resultados obtidos por um modelo, três métricas são amplamente utilizadas: a precisão, a sensibilidade, também conhecida como *recall* ou revocação, e o F-Score. Mas, antes de entender essas métricas, é essencial compreender os seguintes conceitos:

- **Verdadeiro-positivo (VP):** refere-se aos casos em que o modelo identifica corretamente entidades presentes no conjunto de entidades que devem ser identificadas.
- **Verdadeiro-negativo (VN):** refere-se aos casos em que o modelo não identifica entidades que realmente não devem ser identificadas.
- **Falso-positivo (FP):** refere-se aos casos em que o modelo erroneamente identifica entidades que não deveriam ser identificadas.
- **Falso-negativo (FN):** refere-se aos casos em que o modelo falha em identificar entidades que deveriam ser identificadas.

A Figura 2.15 ilustra esses diferentes casos em uma matriz de confusão, onde *gold standard labels* refere-se às saídas esperadas e *system output labels* refere-se às saídas do modelo. Além disso, é possível ver na Figura três possíveis métricas para avaliar o desempenho do modelo: precisão, sensibilidade e acurácia.

Como observado por Jurafsky e Martin (2023), apesar da acurácia parecer ser a métrica ideal, visto que ela indica a porcentagem de casos em que o modelo classificou corretamente as entidades, ela não é a mais adequada quando tem-se textos desbalanceados, ou seja, textos em que a quantidade de entidades a serem reconhecidas e a quantidade de entidades que não devem ser reconhecidas são muito diferentes. Para esses casos, as métricas de precisão e sensibilidade são as mais adequadas.

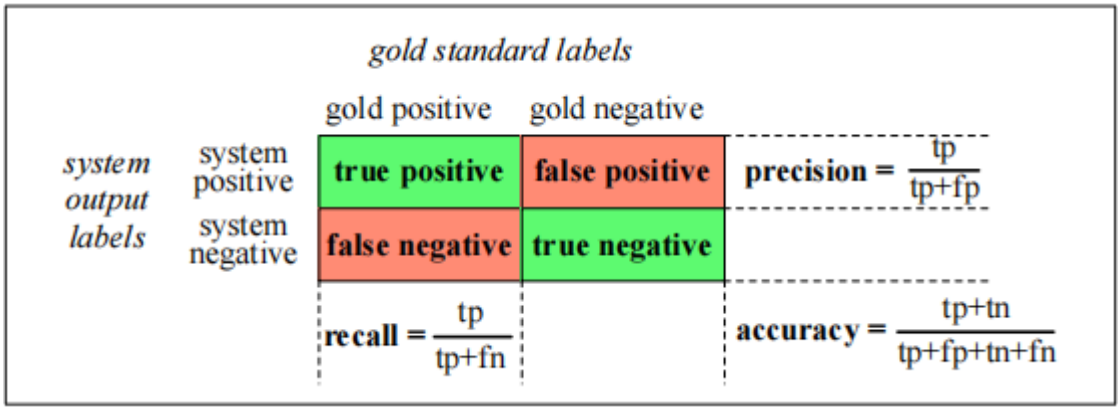


Figura 2.15 – Matriz de confusão para avaliar o desempenho de um modelo de REN.

Fonte: Jurafsky e Martin (2023)

Segundo Jurafsky e Martin (2023), a métrica de precisão indica a porcentagem de casos em que o modelo detectou corretamente como positivo e é calculada por:

$$\text{Precisão} = \frac{VP}{VP + FP}. \tag{2.20}$$

Já a métrica de sensibilidade indica a porcentagem de casos em que o modelo reconheceu entidades que deveriam ser reconhecidas e é calculada pela Equação 2.21:

$$\text{Sensibilidade} = \frac{VP}{VP + FN}. \tag{2.21}$$

Além das métricas de precisão e sensibilidade, que muitas vezes já são o suficiente, existe também o F-Score. Essa medida, introduzida em Van Rijsbergen (1975), incorpora a precisão e a sensibilidade e é calculada pela Equação 2.23:

$$F_\beta = \frac{\left(\beta^2 + 1\right) PR}{\beta^2 P + R}. \tag{2.22}$$

O parâmetro β, como descreve Jurafsky e Martin (2023), pondera a importância da precisão e da sensibilidade. Caso $\beta > 1$, a sensibilidade terá um peso maior, enquanto que para $\beta < 1$ a precisão terá um peso maior. Entretanto, $\beta = 1$ é mais comumente utilizado e indica igual importância das duas medidas. Assim, a Equação 2.23 se torna:

$$F_1 = \frac{2PR}{P + R}. \tag{2.23}$$

Neste trabalho, por padrão, serão utilizadas as medidas de precisão, sensibilidade e F1-Score para avaliar o desempenho dos modelos de REN.

3 Trabalhos Relacionados

Este capítulo tem como objetivo realizar uma revisão da literatura na área de Reconhecimento de Entidades Nomeadas. Para isso, foram levantados trabalhos relacionados a fim de situar este trabalho dentro do contexto científico, além de descobrir quais técnicas estão sendo aplicadas ultimamente.

3.1 Reconhecimento de Entidades Nomeadas em Textos de Publicações do DOU

O Reconhecimento de Entidades Nomeadas tem sua aplicação em diversas áreas. Uma delas é na oferta de seguro garantia para empresas vencedoras de licitações. Segundo a Lei 8.666 em Brasil (1993), os órgãos públicos podem exigir um seguro-garantia que garanta o cumprimento das obrigações por parte das empresas. Por isso, com o propósito de criar um banco de dados de empresas para ofertas de seguro garantia, Albanaz (2020) propôs um modelo de Redes Neurais Convolucionais capaz de reconhecer empresas vencedoras de licitações em publicações de resultados no DOU. A partir do treinamento e teste em uma base de 19.321 publicações, obteve-se como resultado um modelo com 90% de acurácia.

Outro estudo relevante é o de Silva Rodríguez e Bezerra (2020), onde a tarefa de Reconhecimento de Entidades Nomeadas é aplicada na identificação de nomes de agentes públicos em textos jurídicos de atos administrativos. Para isso, foi construída uma base de publicações extraídas do DOU, referentes à portarias de nomeação, exoneração, aposentadoria e outros administrativos. A partir desta base, contendo 10 mil publicações, Silva Rodríguez e Bezerra (2020) utiliza um método baseado em regras para obter 92,9% de acurácia. O ótimo desempenho obtido por um método que não envolve a utilização de modelos de *deep learning* pode ser explicado pela alta padronização das entidades identificadas, que no caso são nomes de agentes públicos, assim como o fato de serem independentes de contexto, permitindo assim a utilização de um método de correspondência direta.

Em Alles, Giozza e Oliveira Alburquerque (2018), o Reconhecimento de Entidades Nomeadas é utilizado para a construção de um *corpus* a partir de textos do DOU. Alles, Giozza e Oliveira Alburquerque (2018) constroem um *corpus* a partir de 470 diários e treinam um modelo de aprendizado supervisionado capaz de identificar as entidades Cargo, Evento, Data, Lei, Lugar, Número, Organização, Pessoa, Processo e Valor-monetário em textos do DOU. Como resultado, foi obtido um modelo com precisão de 95,30%, sensibilidade de 60,70% e F1-Score de 44,50%. Como explicado pelo autor, o baixo F1-Score obtido pode ser atribuído ao fato do trabalho não considerar as complexidades da linguagem natural,

como a ambiguidade das entidades. Além disso, outros fatores podem ter comprometido o desempenho, como a forma de coleta dos textos, que envolveu a conversão de arquivos PDF para JSON, onde se pode ter resultados inconsistentes provenientes de erros de conversão.

3.2 Reconhecimento de Entidades Nomeadas

Apesar de no Capítulo 2 terem sido apresentados diferentes métodos de Reconhecimento de Entidades Nomeadas isoladamente, estes métodos também podem ser utilizados em conjunto com outras técnicas. Em Ferri (2016) é utilizada uma abordagem baseada em dicionário, onde são adicionados mecanismos que visam minimizar o impacto de erros ortográficos no resultado. Nesse sentido, Ferri (2016) utiliza uma busca de entidades por associação aproximada e uma filtragem por métricas de similaridade para comparar os resultados de diferentes combinações destes dois mecanismos. Como melhor resultado, foi obtido uma precisão de 91,84 %, sensibilidade de 96,50% e F1-Score de 88,3%.

Como citado no Capítulo 1, uma das aplicações do Reconhecimento de Entidades Nomeadas é no auxílio à Polícia Federal em investigações de fraudes em licitações públicas. Em Santana (2020) é proposta a aplicação do Reconhecimento de Entidades Nomeadas a fim de auxiliar a Polícia Federal em investigações de lavagem de dinheiro. Para isso, são avaliados os resultados em três cenários: utilização de um modelo pré-treinado em textos genéricos, utilização de um modelo treinado com Relatórios de Inteligência Financeira (RIF) e utilização de um modelo pré-treinado em textos genéricos e treinado com RIFs. Como resultado, observou-se que o refinamento de um modelo pré-treinado apresentou o melhor desempenho, com 69,72% de precisão, 68,74% de sensibilidade e 69,18% de F1-Score.

De forma a avaliar a combinação de textos formais e informais no desempenho de modelos para a tarefa de Reconhecimento de Entidades Nomeadas, Costa (2023) utiliza e compara quatro arquiteturas de modelos (CRF, BiLSTM-CRF, BERT e RoBERTa) para o reconhecimento das entidades Pessoa, Local, Organização, Data, Evento, Fundamento e Produto de Lei em textos relacionados a Projetos de Lei. Duas hipóteses são avaliadas por Costa (2023): se a combinação de textos formais e informais resultam em uma melhora no desempenho dos modelos e se os modelos baseados em *Transformer* possuem um melhor desempenho em textos do domínio legislativo.

Para avaliar a primeira hipótese, Costa (2023) treina os quatro modelos utilizando um *corpus* informal com 285.920 comentários referentes a 6.560 projetos de lei disponibilizados no portal da Câmara dos Deputados, um *corpus* formal com projetos de lei e solicitações de trabalhos e um *corpus* sendo a junção dos dois anteriores. Como resultado, tem-se que a junção de textos formais e informais melhorou a performance dos modelos. Quanto à segunda hipótese, concluiu-se que os modelos baseados em *Transformer* possuem um melhor desempenho, sendo os únicos capazes de compreender as relações semânticas entre o *corpus*

formal e o *corpus* informal, como observado por Costa (2023).

A Tabela 3.5 apresenta, de forma resumida, os objetivos, métodos, tamanho das bases utilizadas para treinamento e teste, e, por fim, os resultados alcançados em cada um dos trabalhos relacionados.

Tabela 3.5 – Resumo dos trabalhos relacionados.

Trabalho	**Objetivo**	**Método**	**Tamanho da base**	**Resultados**
Albanaz (2020)	Reconhecer empresas vencedoras de licitações em publicações de resultados no DOU	Modelo de Redes Neurais Convolucionais	19.321 publicações	Acurácia: 90%
Silva Rodríguez e Bezerra (2020)	Identificar nomes de agentes públicos em textos jurídicos de atos administrativos	Método baseado em regras	10.000 publicações	Acurácia: 92,9%
Alles, Giozza e Oliveira Alburquerque (2018)	Identificar as entidades Cargo, Evento, Data, Lei, Lugar, Número, Organização, Pessoa, Processo e Valor monetário em textos do DOU	Modelo de Aprendizado Supervisionado	470 diários	Precisão: 95,30% Sensibilidade: 60,70% F1-Score: 44,50%

Tabela 3.5 – Resumo dos trabalhos relacionados.

Trabalho	**Objetivo**	**Método**	**Tamanho da base**	**Resultados**
Ferri (2016)	Comparar os resultados de diferentes combinações entre dois mecanismos: busca de entidades por associação aproximada e filtragem por métricas de similaridade	Método baseado em dicionário	Informação não encontrada	Precisão: 91,84% Sensibilidade: 96,50% F1-Score: 88,30%
Santana (2020)	Reconhecer entidades nomeadas em Relatórios de Inteligência Financeira (RIF)	Modelo de Aprendizado Supervisionado	35 RIFs	Precisão: 69,72% Sensibilidade: 68,74% F1-Score: 69,18%
Costa (2023)	Reconhecer as entidades Pessoa, Local, Organização, Data, Evento, Fundamento e Produto de Lei em textos relacionados a Projetos de Lei	CRF, BiLSTM-CRF, BERT e RoBERTa	85.920 comentários referentes a 6.560 projetos de lei	BERT (melhor modelo): Precisão: 77,69% Sensibilidade: 81,64% F1-Score: 78,65%

4 Metodologia

Com o objetivo de estabelecer uma base de publicações anotadas e conduzir o treinamento dos modelos de *Transformers* para o reconhecimento de entidades nomeadas em publicações de convênios, este projeto foi dividido em cinco grandes etapas, como ilustrado na Figura 4.16. Nessa figura são apresentados resumidamente a ação, o objetivo e os resultados esperados em cada etapa. A seguir, serão fornecidos detalhes sobre cada uma dessas etapas.

Figura 4.16 – Etapas da metodologia proposta.

4.1 Mapeamento das Entidades

A primeira etapa consistiu no mapeamento das possíveis entidades de um convênio, visando avaliar a importância de cada uma delas com base em alguns aspectos. Essa etapa é de suma importância, uma vez que um mapeamento eficiente tornará as etapas subsequentes de busca e validação das publicações mais ágeis e eficazes.

Para realizar o mapeamento, utilizou-se como base um arquivo CSV disponibilizado em Portal da Transparência (2023b), que contém os dados de 27 entidades para cada convênio criado a partir de 1996. A periodicidade de atualização deste arquivo é semanal, e para este trabalho, utilizou-se a versão com os dados de convênios criados até 28/10/2022.

Conforme mencionado anteriormente, foram mapeadas 27 entidades de convênios, listadas na Tabela 4.6 extraída de Portal da Transparência (2023c). No processo de mapea-

mento dessas entidades, considerou-se uma amostra de 50 convênios, nos quais dois aspectos de extrema importância foram levados em consideração: a unicidade e a frequência das entidades nas publicações. Esses dois aspectos foram desenvolvidos neste trabalho a fim de tornar a busca de publicações mais eficiente e solucionar problemas ocorridos durante a etapa de busca, os quais serão detalhados mais adiante.

4.1.1 Unicidade

A unicidade é uma característica crucial para as entidades que serão utilizadas na busca de publicações relacionadas a um convênio específico. Além de limitar a quantidade de publicações retornadas, a busca por meio de uma entidade que possui unicidade é altamente eficiente, pois a maioria das publicações retornadas está diretamente relacionada ao convênio em questão.

Por outro lado, a busca de publicações por meio de uma entidade sem a característica de unicidade resultaria em um grande número de publicações que compartilham essa entidade, mas que não estão relacionadas ao convênio de interesse. Um exemplo claro de uma entidade com unicidade é o Número do Convênio, já que cada convênio possui um número exclusivo. Por outro lado, a Unidade Federativa (UF) do convênio não possui unicidade, uma vez que é compartilhada por vários convênios e instrumentos públicos.

Portanto, para mapear as entidades com base na característica de unicidade, o seguinte processo foi realizado na amostra de 50 convênios:

1. Para cada entidade de cada convênio, realizou-se a busca por publicações em Imprensa Nacional (2023), utilizando a entidade em questão como campo de pesquisa;
2. Posteriormente, para cada publicação encontrada, verificou-se se esta estava relacionada ao convênio em questão. No caso das entidades que retornaram um grande número de publicações, em torno de milhares, considerou-se que o nível de unicidade era baixo, uma vez que um convênio gera dezenas de publicações em média;
3. Com base no total de publicações retornadas e na quantidade de publicações relacionadas ao convênio, calculou-se a porcentagem de publicações retornadas que estavam realmente relacionadas ao convênio. Esse valor serviu como referência para determinar o nível de unicidade de cada entidade.

4.1.2 Frequência

A frequência das entidades nas publicações é outro aspecto importante a ser considerado. Essa característica não se refere à frequência com que uma entidade aparece em uma única publicação, mas sim à frequência com que ela ocorre em diferentes publicações de um convênio. Quanto mais frequente for a presença da entidade nas publicações, mais eficaz ela

será para a busca das mesmas, garantindo que as publicações relacionadas ao convênio não sejam perdidas.

Para mapear as entidades em relação à característica de frequência, o seguinte processo foi realizado na amostra de 50 convênios:

1. Para cada entidade de cada convênio, foi realizada uma busca por publicações em Imprensa Nacional (2023), utilizando a entidade específica como campo de pesquisa;
2. Posteriormente, para cada publicação encontrada, verificou-se se esta estava relacionada ao convênio em questão. Nesse processo, foram desconsideradas as entidades que geraram um número excessivamente alto de publicações, alcançando a faixa de milhares;
3. Após obter todas as publicações relacionadas ao convênio, calculou-se a porcentagem de publicações que mencionavam a entidade em seu texto. Esse cálculo foi usado como referência para determinar o nível de frequência de cada entidade.

Conforme mencionado previamente, o mapeamento e a seleção adequada das entidades com base nas características de unicidade e frequência são de extrema importância. Se uma entidade possui unicidade, mas não é frequentemente mencionada nas publicações, ela terá pouca relevância na busca por essas publicações. Da mesma forma, se uma entidade é frequente, mas não possui unicidade, a busca por publicações se torna ineficiente e, em alguns casos, inviável. Portanto, a seleção das entidades que serão utilizadas na busca e validação das publicações deve buscar um equilíbrio entre essas duas características.

Tabela 4.6 – Descrição das entidades de um convênio.

ENTIDADE	**DESCRIÇÃO**
Número Convênio	Número que identifica o convênio
UF	Sigla da Unidade Federativa do convenente
Código SIAFI Município	Código, no SIAFI (Sistema Integrado de Administração Financeira), do município do convenente
Nome Município	Nome do município do convenente
Situação Convênio	Situação em que se encontra o convênio
Número Original	Número Original do convênio
Número Processo do Convênio	Número do processo do convênio
Objeto do Convênio	Aquilo pactuado entre o Governo Federal concedente e o convenente beneficiado no município
Código Órgão Superior	Código do Órgão Superior concedente ÓRGÃO SUPERIOR - Unidade da Administração Direta que tenha entidades por ele supervisionadas.

Tabela 4.6 – Descrição das entidades de um convênio.

ENTIDADE	**DESCRIÇÃO**
	Fonte: Manual do SIAFI
Nome Órgão Superior	Nome do Órgão Superior concedente
Código Órgão Concedente	Órgão Concedente: Órgão da administração pública federal direta, autárquica ou fundacional, empresa pública ou sociedade de economia mista, responsável pela transferência dos recursos financeiros ou pela descentralização dos créditos orçamentários destinados à execução do objeto do convênio.
Nome Órgão Concedente	Nome do Órgão concedente
Código UG Concedente	Código da UG concedente
Nome UG Concedente	Nome da UG concedente
Código Convenente	Convenente: Órgão da administração direta, autárquica ou fundacional, empresa pública ou sociedade de economia mista, de qualquer esfera de governo, ou organização particular com a qual a administração federal pactua a execução de programa, projeto ou atividade, ou evento mediante a celebração de convênio. É quem recebe os recursos do Governo Federal.
Tipo Convenente	Tipo do convenente
Nome Convenente	Nome do convenente
Tipo Ente Convenente	Municipal ou Estadual
Tipo Instrumento	Tipo de instrumento
Valor Convênio	Valor do Convênio: É o valor correspondente à participação do concedente. É adicionado ao valor original do convênio a parcela (999) que corresponde a rendimento de aplicação financeira, quando for o caso.
Valor Liberado	Valor Liberado (convênio): Valor total liberado pelo Governo Federal até a data de atualização da base de dados. É adicionado ao valor original do convênio a parcela (999) que corresponde a rendimento de aplicação financeira, quando for o caso.
Data Publicação	Data de Publicação do convênio

Tabela 4.6 – Descrição das entidades de um convênio.

ENTIDADE	DESCRIÇÃO
Data Início Vigência	Data de início da vigência do convênio
Data Fim Vigência	Data de fim da vigência do convênio
Valor Contrapartida	Valor da Contrapartida (convênio): Valor correspondente à participação do convenente no convênio, para a execução do objeto.
Data Última Liberação	Data em que foi feita a última liberação de recursos pelo concedente ao convenente.
Valor Última Liberação	Valor Última Liberação (convênio): Valor relativo à última liberação de recursos do convênio pelo concedente ao convenente. .

4.1.3 Variações das entidades

Além de mapear as entidades em termos de unicidade e frequência, a etapa de mapeamento também foi crucial para compreender as diferentes formas pelas quais uma mesma entidade é mencionada nos textos das publicações. Por exemplo, sabe-se que o formato de um CNPJ é XX.XXX.XXX/YYYY-ZZ, em que X, Y e Z são dígitos de 0 a 9. No entanto, nos textos das publicações, ele pode aparecer no formato XXXXXXXXYYYYZZ, XX.XXX.XXX/YYYY ou XXXXXXXXYYYY. Portanto, para cada entidade, foram definidas variações considerando a inspeção manual de centenas de publicações, o que se mostrou suficiente para esse propósito.

4.2 Busca das Publicações

Com base nas entidades selecionadas na etapa anterior e em seus respectivos valores para cada convênio, extraídos de Portal da Transparência (2023b), esta etapa buscou e armazenou as publicações do portal do DOU em Imprensa Nacional (2023). Conforme ilustrado no fluxograma da Figura 4.17, esta etapa consistiu em quatro laços de repetição aninhados. O primeiro laço percorreu a lista de convênios, o segundo percorreu a lista de entidades para cada convênio, o terceiro percorreu a lista de variações para cada entidade de cada convênio, e o quarto percorreu a lista de publicações retornadas na busca.

O processo de busca das publicações em Imprensa Nacional (2023) foi realizado por meio da técnica de raspagem de dados, ou *web scraping*, que envolve a extração de informações de uma página *web* ou sistema específico, conforme descrito em Netrin (2023). Esse processo foi dividido em três fases, que serão detalhadas a seguir.

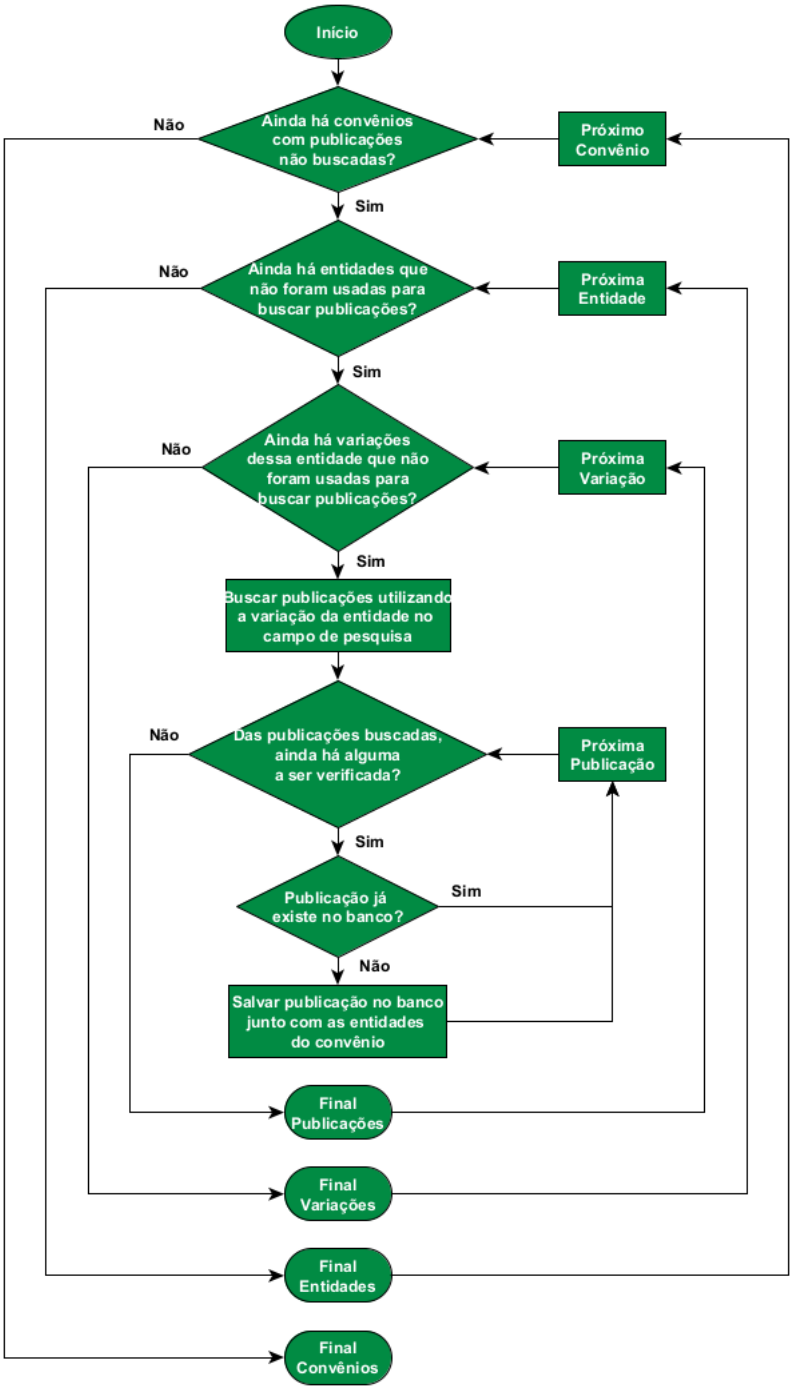


Figura 4.17 – Fluxograma de execução da etapa de busca das publicações.

4.2.1 Montagem da URL

Para realizar a pesquisa de publicações no portal do DOU foi necessário um processo de montagem da URL da requisição HTTP. Com base no conteúdo inserido no campo de pesquisa e nos filtros de busca mostrados na Figura 4.18, o portal do DOU gera uma cadeia de consulta, ou *query string*, que é anexada à URL base do portal do DOU. No exemplo apresentado na Figura 4.18, a *query string* gerada é *"dou?q="+exemplo"&s=do3&exactDate=personalizado &sortType=0&publishFrom=01-01-2018&publishTo=05-06-2023"*. Portanto, a cada busca de

publicações, a URL de requisição foi gerada com base na variação da entidade a ser utilizada no campo de pesquisa, bem como nos filtros pré-definidos, detalhados na Tabela 4.7.

Figura 4.18 – Filtros de pesquisa do portal do DOU.

Tabela 4.7 – Filtros de pesquisa utilizados.

Filtro	**Opção escolhida**	**Justificativa**
Tipo de pesquisa	Resultado exato	Retorna resultados mais precisos
Forma de pesquisa	Pesquisa ato-a-ato	Forma de pesquisa em que as publicações são mostradas como páginas web, possibilitando a utilização da técnica de *web scraping*
Onde pesquisar	Tudo	Entidade pode estar no título ou no texto
Ordenação	Por data	Este filtro não é relevante para os objetivos deste trabalho
Data	Personalizado: 01/01/2018 até a data da pesquisa	Apenas as publicações a partir de 01/01/2018 estão presentes na forma de pesquisa ato-a-ato
Jornal	Seção 3	Seção em que estão as publicações relativas a convênios.

4.2.2 Busca das Publicações

Após obter a URL de requisição, gerada a partir de uma entidade utilizada no campo de pesquisa e dos filtros de pesquisa, foi feita uma requisição HTTP para obter a página da

web resultante da pesquisa, no formato HTML. Por meio da técnica de *web scraping*, as URLs de cada uma das publicações retornadas na pesquisa foram extraídas.

4.2.3 Extração do Texto das Publicações

Por meio da lista de URLs das publicações encontradas, foi possível realizar uma requisição HTTP para cada URL, a fim de obter as páginas da *web* em formato HTML das respectivas publicações. Novamente, por meio da técnica de *web scraping*, os textos das publicações foram extraídos e armazenados, mantendo a associação com o convênio correspondente. A Figura 4.19 ilustra um exemplo de consulta por publicações no portal do DOU, bem como um exemplo de publicação no formato *web*, disponível apenas para publicações a partir de 2018.

Figura 4.19 – Imagens do portal de consulta de publicações do DOU.

4.3 Validação das Publicações

Durante a etapa de busca das publicações é possível que sejam retornadas algumas que não possuam relação com o convênio, mesmo após o mapeamento das entidades. Essas publicações podem impactar o desempenho do modelo, uma vez que podem resultar em anotações incorretas. Por isso, foi adicionada uma etapa de validação das publicações, após a busca, para garantir que elas realmente pertençam ao convênio em questão.

O fluxograma desta etapa pode ser observado na Figura 4.20. Nessa etapa são utilizadas as entidades mapeadas como frequentes nas publicações, mas que não foram utilizadas na etapa de busca, para validar as publicações. Para cada uma dessas entidades, verifica-se se elas estão presentes na publicação e incrementa-se um contador em caso afirmativo. Se

houver a presença de pelo menos três entidades, considera-se que a publicação pertence ao convênio; caso contrário, ela é descartada. O valor mínimo de três entidades foi definido empiricamente com base na etapa de mapeamento das entidades. Para um valor mínimo de duas entidades, por exemplo, viu-se que o problema das publicações não relacionadas ao convênio persistia para a maioria dos convênios, enquanto que, para um valor mínimo de quatro entidades, viu-se que publicações relacionadas ao convênio eram descartadas. Logo, o valor mínimo de três entidades foi considerado como um bom meio termo entre as duas situações.

4.4 Anotação das Publicações

Ao treinar um modelo de *Machine Learning* é crucial fornecer uma quantidade substancial de dados. No entanto, esses dados precisam ser preparados de maneira que o modelo seja capaz de identificar padrões e inferências com facilidade. Isso geralmente envolve a adição de metadados ao conjunto de dados, conforme explicado em Pustejovsky e Stubbs (2012).

Em PLN, esses metadados aparecem como forma de marcações que destacam informações relevantes no texto, sendo cada marcação chamada de anotação sobre o texto de entrada. Ainda segundo Pustejovsky e Stubbs (2012), para garantir um aprendizado eficiente e efetivo do modelo, as anotações dos dados devem ser acuradas e relevantes para os objetivos das predições. Portanto, todas as etapas concluídas até o momento - mapeamento das entidades, busca e validação das publicações - têm como propósito aprimorar a acurácia das anotações e extrair o melhor desempenho dos modelos.

Nesta etapa, todas as publicações buscadas e validadas passaram pelo processo de anotação e compuseram a base de publicações anotadas, que posteriormente foi utilizada para o treinamento do modelo. Para auxiliar na anotação das publicações, foi empregada a ferramenta PhraseMatcher, documentada em spaCy (2023c), que permite a identificação de grandes listas de terminologias em um texto por meio de correspondencia direta. A Figura 4.21 apresenta um exemplo visual de uma publicação anotada, no caso, um extrato de convênio. As partes marcadas no texto correspondem a diferentes entidades, sendo que cada cor representa uma entidade distinta, cujo rótulo é destacado em negrito. A produção desse exemplo visual foi realizada com o auxílio da ferramenta mencionada em spaCy (2023a).

4.5 Treinamento

Após a construção da base de publicações anotadas, a última etapa consistiu no treinamento dos modelos de *Transformers*. Para essa finalidade, foi empregada a ferramenta em Honnibal et al. (2020), devido à sua facilidade de integração com os dados anotados e

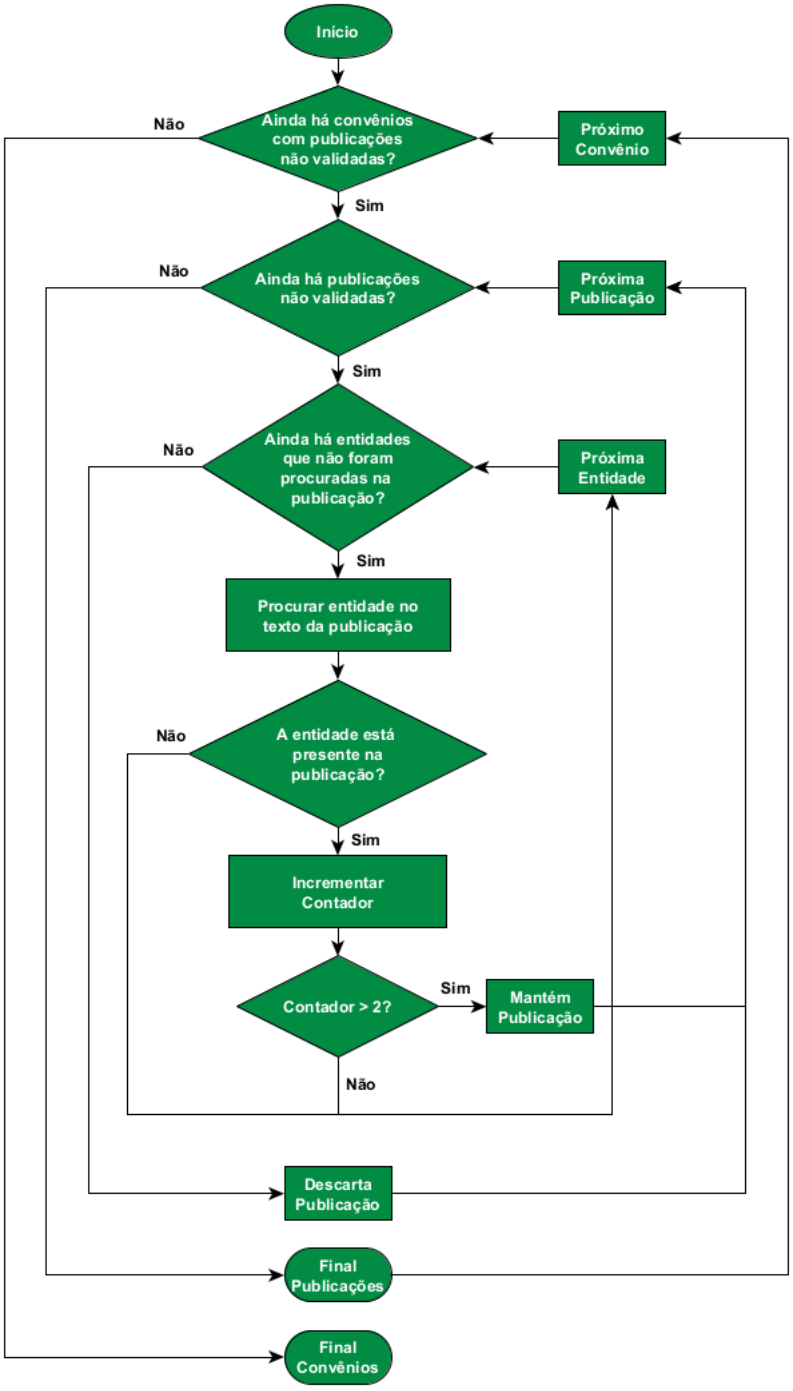


Figura 4.20 – Fluxograma de execução da etapa de validação das publicações.

também com a biblioteca de modelos em Wolf et al. (2020). Nessa etapa, foram utilizados e comparados dois modelos pré-treinados, os quais serão detalhados a seguir.

4.5.1 BERT

O modelo BERT, introduzido em Devlin et al. (2019), foi o pioneiro entre os modelos de *Transformers* a utilizar um método que possibilita que o modelo veja textos inteiros de

Figura 4.21 – Exemplo de uma publicação anotada.

uma vez, incluindo o contexto à esquerda e à direita, como ilustrado na Figura 4.22. Esse método consiste de um codificador de transformador bidirecional treinado por meio de modelagem de linguagem mascarada. Esse modelo permitiu com que se atingisse o estado da arte em tarefas de PLN complexas, nas quais é crucial levar em consideração o contexto global do texto.

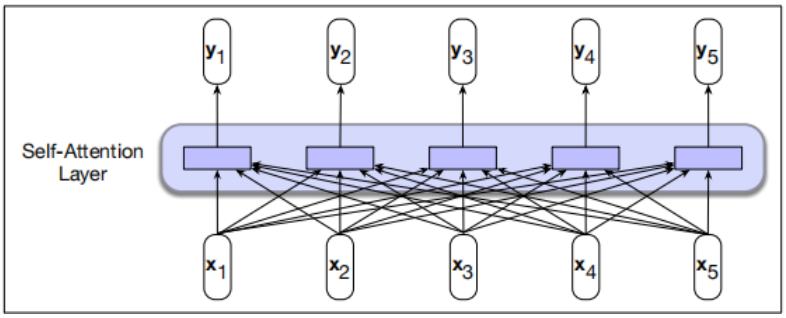


Figura 4.22 – Fluxo de informação em um modelo de autoatenção bidirecional.

Fonte: Jurafsky e Martin (2023)

Neste trabalho utilizou-se uma variante do modelo BERT denominada BERTimbau. Essa variante, introduzida em Fábio Souza, Nogueira e Lotufo (2020), consiste em um modelo BERT pré-treinado para o português do Brasil que alcança o estado da arte em três tarefas de PLN: Reconhecimento de Entidade Nomeada, Similaridade Textual de Sentença e Reconhecimento de Vinculação Textual.

4.5.2 RoBERTa

O modelo RoBERTa, introduzido em Liu et al. (2019), apresenta a mesma arquitetura do modelo BERT, porém com aprimoramentos no pré-treinamento. De acordo com Liu et al. (2019), o modelo BERT foi subtreinado e poderia atingir ou superar o desempenho de todos os modelos subsequentes. Conforme descrito em Hugging Face (2023), algumas otimizações implementadas no pré-treinamento do modelo BERT pelo RoBERTa foram:

- Mascaramento dinâmico: os *tokens* são mascarados de forma diferente em cada época, enquanto o BERT faz isso de uma vez por todas;

- Treinamento com lotes maiores;
- Utilização do BPE (do inglês *Byte-Pair Encoding*) com bytes como uma subunidade e não caracteres (devido a caracteres *unicode*).

Neste trabalho foi empregado o XLM-RoBERTa, proposto em Conneau et al. (2020), um grande modelo de linguagem multilíngue e que possui a mesma implementação do modelo RoBERTa. O XLM-RoBERTa é pré-treinado em 100 idiomas distintos e possui a capacidade de identificar o idioma correto.

5 Resultados

Neste capítulo serão apresentados os resultados deste estudo, divididos em três seções. A primeira seção abordará os resultados obtidos na etapa de mapeamento das entidades de convênios. Na segunda seção serão discutidos os resultados das etapas de busca, validação e anotação das publicações, que compõem a construção da base de dados anotada. Por fim, a terceira seção apresentará os resultados do treinamento dos dois modelos de *Transformer* e realizará uma comparação entre eles.

5.1 Mapeamento das Entidades

Os resultados do estudo das entidades de convênios são apresentados na Tabela 5.8. Como comentado no capítulo anterior, foram mapeadas as características de unicidade e frequência de cada entidade na busca por publicações. Para facilitar a visualização, os resultados do mapeamento das entidades foram classificados em três níveis: baixo, médio e alto. O nível baixo indica que a entidade possui pouca ou nenhuma unicidade ou frequência, enquanto o nível alto indica que a entidade possui alta unicidade ou frequência.

Tabela 5.8 – Resultado do mapeamento das entidades.

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**

Tabela 5.8 – Resultado do mapeamento das entidades.

Entidade	Unicidade	Frequência
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

A principal característica que influencia o desempenho da etapa de busca de publicações é a unicidade. É importante que a entidade tenha uma unicidade média ou alta para ser considerada na busca de publicações, visto que uma unicidade baixa resultaria em um número muito grande de publicações não relacionadas ao convênio, e tornaria a busca demorada e ineficiente. Portanto, para a busca de publicações foram selecionadas apenas as entidades que obtiveram uma unicidade média ou alta:

- Número Convênio;
- Número Original;
- Número Processo do Convênio;
- Objeto do Convênio;
- Código Convenente.

Em relação à validação das publicações, a frequência é a principal característica a ser considerada, uma vez que uma entidade sem essa característica apenas consumiria tempo e recursos. Para garantir a eficiência da validação foram selecionadas apenas as entidades que obtiveram uma frequência média ou alta, e que não foram utilizadas na busca de publicações:

- UF;
- Nome Município;
- Nome Órgão Superior;
- Nome Órgão Concedente;

- Código UG Concedente;
- Nome Convenente;
- Tipo Ente Convenente;
- Valor Convênio;
- Valor Liberado;
- Data Início Vigência;
- Data Fim Vigência;
- Valor Contrapartida;
- Valor Última Liberação.

As entidades Código SIAFI Município, Situação Convênio, Código Órgão Superior, Código Órgão Concedente, Nome UG Concedente, Tipo Convenente, Tipo Instrumento, Data Publicação e Data Última Liberação não foram incluídas nas etapas de busca e validação de publicações devido à sua baixa unicidade e frequência. Apesar de não serem adequadas à coleta das publicações, essas entidades ainda contêm informações relevantes sobre o convênio e foram submetidas ao processo de anotação normalmente.

5.2 Base de Publicações Anotadas

Ao final das etapas de busca, validação e anotação das publicações, foi construída uma base de publicações anotadas. No total foram anotadas 192.900 publicações provenientes de 71.287 convênios. Em média, cada convênio teve 2,7 publicações anotadas. A Tabela 5.9 apresenta a quantidade de anotações para cada entidade. É importante ressaltar que esses números não representam a quantidade de convênios ou publicações em que a entidade foi anotada, mas sim a quantidade de anotações da entidade. Ou seja, se a entidade foi citada mais de uma vez em uma mesma publicação, todas as ocorrências foram contabilizadas.

A partir da Tabela 5.9, pode-se inferir que as entidades com o maior número de anotações foram: UF, Nome Órgão Superior, Tipo Ente Convenente, Data Início Vigência e Valor Última Liberação. Por outro lado, as entidades menos anotadas foram: Código Órgão Superior, Nome UG Concedente, Tipo Convenente, Valor Liberado e Data Publicação. Esses números comprovam os resultados da Tabela 5.8 e corroboram com a etapa de mapeamento das entidades.

O alto número de anotações para algumas entidades pode ser explicado pelo fato de que informações do convênio podem estar presentes tanto em publicações específicas e

pequenas, referentes a um único convênio, quanto em publicações maiores que abrangem informações de diversos convênios. Nessas publicações maiores é comum que os convênios compartilhem o mesmo valor para essas entidades, as quais geralmente não possuem unicidade. Por exemplo, pode haver uma publicação que abranja vários convênios da mesma UF, fazendo com que essa entidade seja contabilizada várias vezes na publicação, mesmo estando no escopo de outros convênios.

Tabela 5.9 – Quantidade de anotações para cada entidade.

Entidade	**Quantidade de Anotações**
Número Convênio	92.891
UF	1.361.105
Código SIAFI Município	23.166
Nome Município	68.081
Situação Convênio	4.004
Número Original	25.771
Número Processo do Convênio	38.371
Objeto do Convênio	114.439
Código Órgão Superior	1.114
Nome Órgão Superior	1.495.540
Código Órgão Concedente	17.774
Nome Órgão Concedente	66.591
Código UG Concedente	301.681
Nome UG Concedente	1.383
Código Convenente	102.837
Tipo Convenente	191
Nome Convenente	149.594
Tipo Ente Convenente	2.274.833
Tipo Instrumento	3.059
Valor Convênio	171.283
Valor Liberado	162
Data Publicação	1
Data Início Vigência	2.190.278
Data Fim Vigência	508.623
Valor Contrapartida	137.369
Data Última Liberação	8.763
Valor Última Liberação	1.275.212

5.3 Treinamento

Nesta seção serão apresentados brevemente os resultados do treinamento de cada um dos dois modelos e, em seguida, será realizado um comparativo entre eles. Para o treinamento dos modelos foi utilizado um servidor disponibilizado pelo Departamento de Ciência da Computação da Universidade de Brasília. O servidor possui duas placas de processamento gráfico Tesla V100S-PCIe, cada uma com 32 GB de memória, nas quais foram realizados os treinamentos.

5.3.1 BERTimbau

Os resultados do treinamento do modelo BERTimbau podem ser vistos na Tabela 5.11. É possível ver que o modelo obteve um ótimo desempenho geral, apresentando um *F1-Score* igual a 0,94, uma precisão igual a 0,944 e uma sensibilidade, ou *recall*, igual a 0,935. Portanto, esses resultados mostram que o modelo atinge um ótimo desempenho para o Reconhecimento de Entidades Nomeadas.

Ao observar o desempenho por entidade, pode-se notar que há uma grande variação, como já era esperado. Algumas entidades obtiveram um *F1-Score* alto, como Código Órgão Superior, Nome Órgão Superior e Código Convenente, enquanto que outras obtiveram um *F1-Score* zerado ou baixo, como Código SIAFI Município, Situação Convênio e Objeto do Convênio. Considerando que 75% das entidades tiveram um *F1-Score* acima de 0,8, é possível concluir que o modelo obteve um ótimo desempenho por entidade.

A grande diferença de *F1-Score* entre as entidades pode ser explicada, entre outros aspectos, pela diferença na quantidade de anotações. Como pode ser visto na Figura 5.23, as entidades com maior número de anotações tiveram, em geral, um *F1-Score* maior. Isso pode ser melhor observado pela linha de tendência no gráfico, apesar de não se poder garantir que há essa tendência na prática.

Um caso interessante a se notar é o do Objeto do Convênio, visto que foi o terceiro pior desempenho, mesmo tendo uma quantidade considerável de anotações, cerca de 114 mil. Isso pode ser explicado pela característica peculiar dessa entidade. Como observado na Figura 5.24, o Objeto do Convênio pode aparecer nas publicações como uma frase extensa, o que dificulta o aprendizado do modelo.

5.3.2 XLM-RoBERTa

Como pode ser visto na Tabela 5.12, o modelo XLM-RoBERTa obteve um ótimo desempenho geral, apresentando um *F1-Score* igual a 0,942, uma precisão igual a 0,949 e uma sensibilidade, ou *recall*, igual a 0,935. Em relação ao desempenho por entidade, o modelo também teve um ótimo desempenho, onde 75% das entidades tiveram um *F1-Score* maior

que 0,8. Dentre as entidades com o maior *F1-Score*, estão o Código Órgão Superior, Código Convenente e Nome Órgão Superior. Por outro lado, as entidades com menor *F1-Score* foram Código SIAFI Município, Situação Convênio e Tipo Instrumento.

A Figura 5.25 mostra uma tendência de aumento do *F1-Score* a medida em que a quantidade de anotações aumenta, assim como no modelo BERTimbau. Essa tendência é natural visto que quanto mais exemplos de uma entidade o modelo possui, melhor será o seu aprendizado sobre aquela entidade. Apesar disso, ainda sim algumas entidades tiveram um *F1-Score* alto com uma quantidade de anotações relativamente baixa. Esse é o caso das entidades Código Órgão Superior, Nome UG Concedente e Código Órgão Concedente, por exemplo. Isso pode ser atribuído ao fato de que essas entidades aparecem, geralmente, da mesma forma nos textos das publicações.

As entidades Tipo Convenente, Valor Liberado e Data Publicação não tiveram resultados, tanto no modelo BERTimbau quanto no XLM-RoBERTa, devido à baixa quantidade de anotações.

5.3.3 Comparativo

No desempenho geral, os dois modelos tiveram um resultado muito parecido, com o XLM-RoBERTa sendo ligeiramente superior. Já em relação ao desempenho por entidade, o gráfico da Figura 5.26 mostra que o modelo XLM-RoBERTa teve um desempenho ligeiramente inferior nas entidades com quantidade de anotações baixa e ligeiramente superior

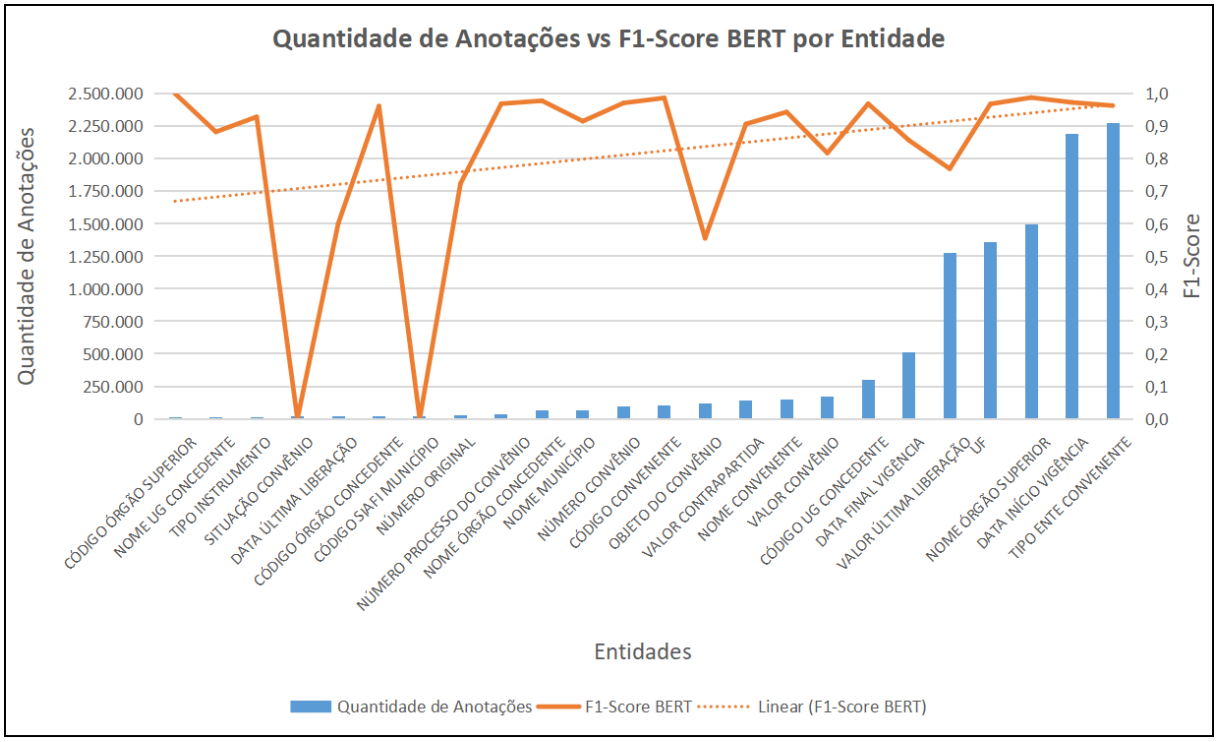


Figura 5.23 – Quantidade de anotações vs *F1-Score* por entidade para o modelo BERT.

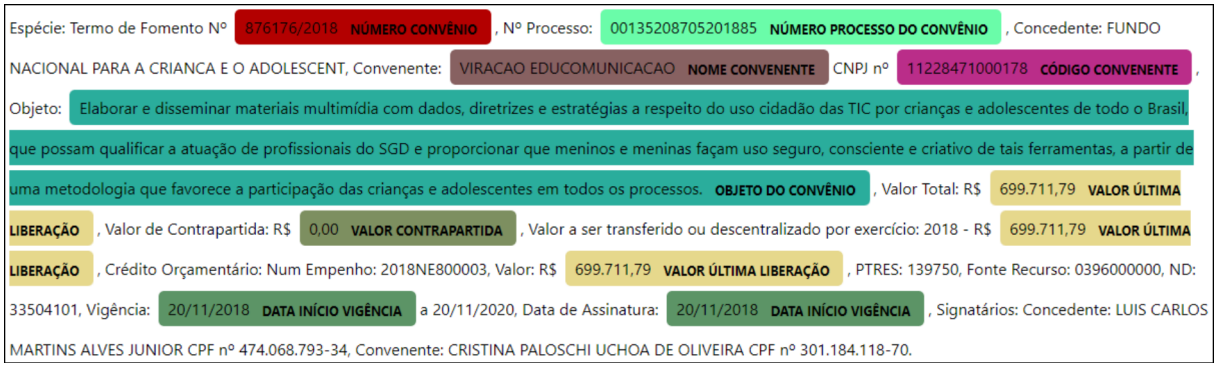


Figura 5.24 – Publicação com objeto do convênio extenso.

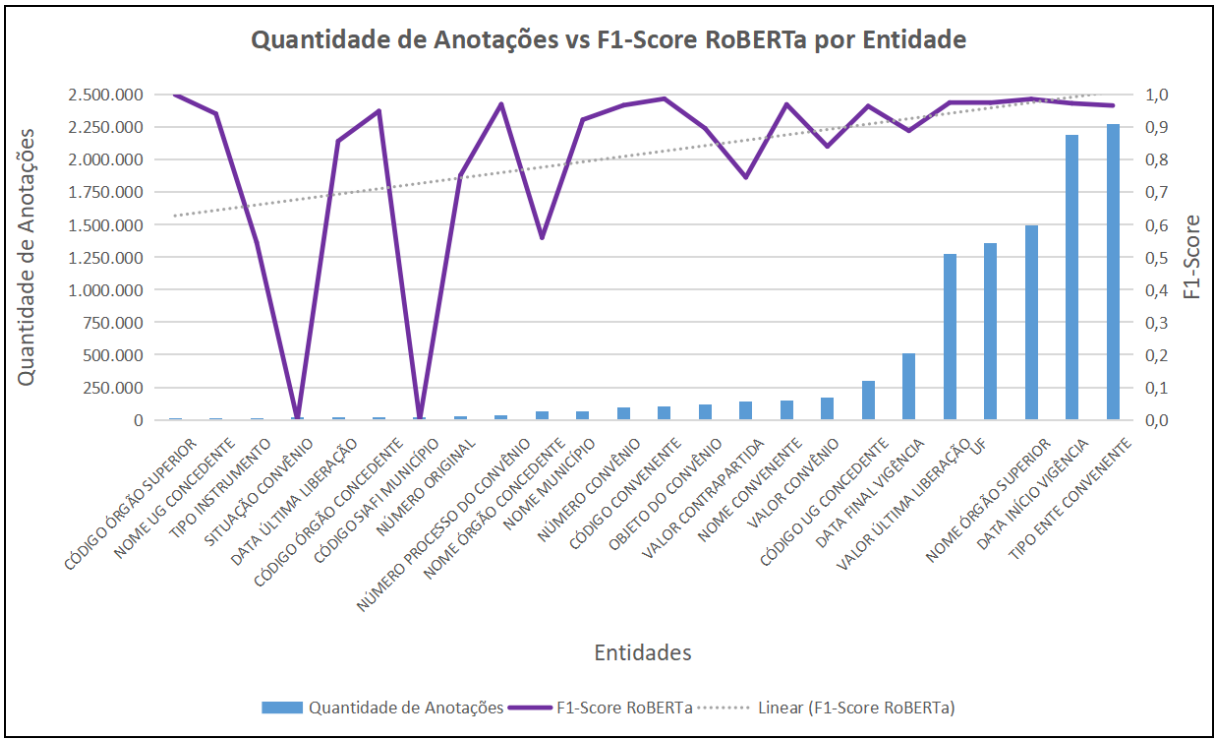


Figura 5.25 – Quantidade de anotações vs *F1-Score* por entidade para o modelo RoBERTa.

nas entidades com quantidade de anotações alta. No geral, o formato das duas curvas se mantiveram parecidos.

O contraste entre os dois modelos fica claro na Tabela 5.10. Ao tomar os resultados do modelo BERTimbau como referência, é possível observar que o modelo XLM-RoBERTa teve um desempenho significativamente melhor para as entidades Objeto do Convênio, Data Última Liberação e Valor Última Liberação. Por outro lado, houve uma queda de desempenho em relação às entidades Nome Órgão Concedente, Tipo Instrumento e Valor Contrapartida. Com relação às outras entidades que tiveram alguma variação, o modelo XLM-RoBERTa obteve um desempenho melhor em aproximadamente 73% delas.

O aumento expressivo do *F1-Score* no reconhecimento da entidade Objeto do Con-

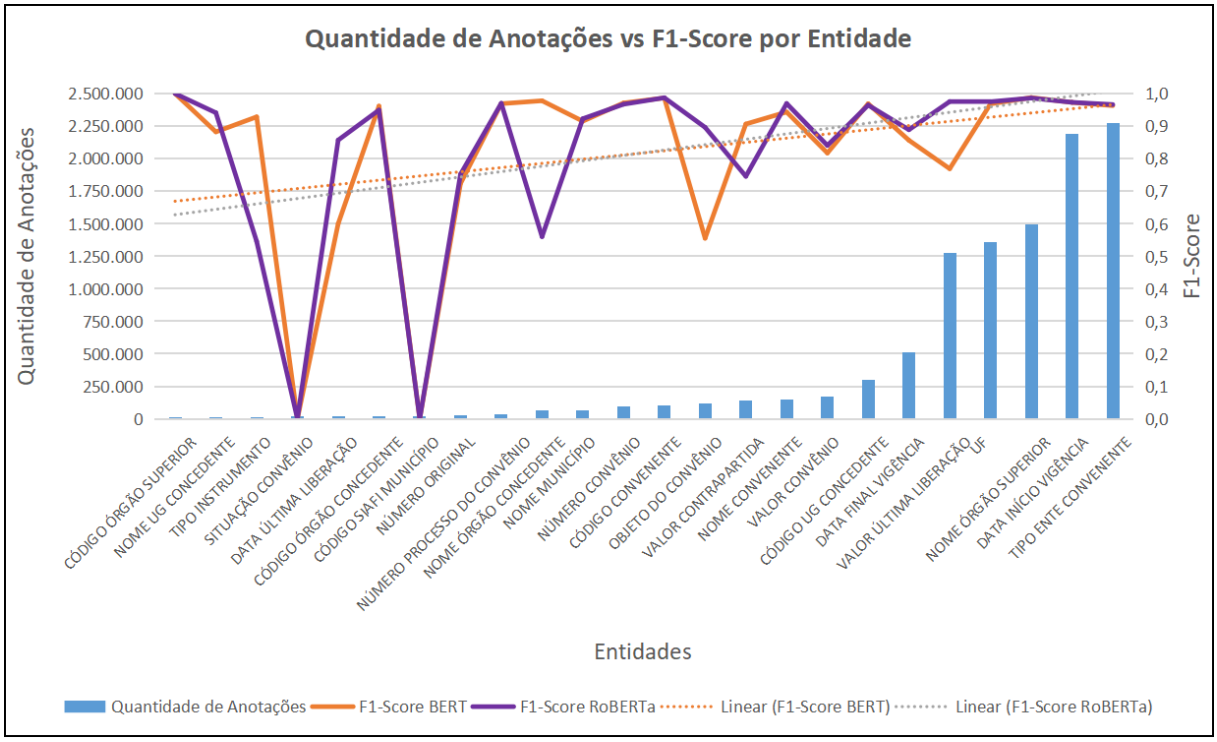


Figura 5.26 – Quantidade de anotações vs *F1-Score* por entidade para os dois modelos.

vênio evidencia o melhor desempenho do modelo XLM-RoBERTa no reconhecimento de entidades complexas. Apesar de todas as entidades terem sua importância no contexto de investigação de fraudes, o reconhecimento dessas entidades complexas é fundamental, pois muitas vezes são entidades que dependem do contexto, em que métodos mais simples, como o baseado em regras, não seriam capazes de reconhecer. Em contrapartida, o modelo XLM-RoBERTa obteve um pior desempenho no reconhecimento da entidade Nome Órgão Concedente, que é uma entidade relevante para a vinculação entre publicações e identificação de partes de uma investigação.

Tabela 5.10 – Comparativo de *F1-Score* por entidade.

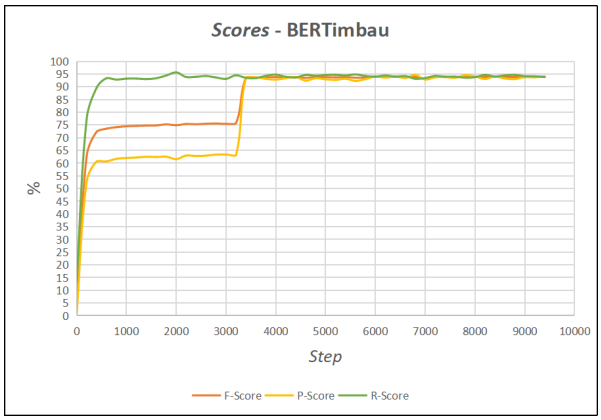
Entidade	***F1-Score***		
	BERTimabu	**XLM-RoBERTa**	**XLM-RoBERTa - BERTimbau [%]**
CÓDIGO ÓRGÃO SUPERIOR	1,000	1,000	0,000
NOME UG CONCEDENTE	0,882	0,941	6,667
TIPO INSTRUMENTO	0,929	0,545	-41,259
SITUAÇÃO CONVÊNIO	0,000	0,000	0,000
DATA ÚLTIMA LIBERAÇÃO	0,600	0,857	42,857
CÓDIGO ÓRGÃO CONCEDENTE	0,962	0,949	-1,266
CÓDIGO SIAFI MUNICÍPIO	0,000	0,000	0,000

Tabela 5.10 – Comparativo de *F1-Score* por entidade.

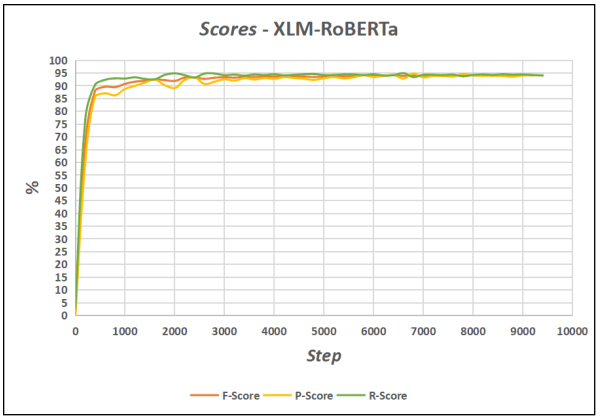
Entidade	***F1-Score***		
	BERTimabu	**XLM-RoBERTa**	**XLM-RoBERTa - BERTimbau [%]**
NÚMERO ORIGINAL	0,723	0,752	3,938
NÚMERO PROCESSO DO CONVÊNIO	0,969	0,970	0,127
NOME ÓRGÃO CONCEDENTE	0,978	0,560	-42,717
NOME MUNICÍPIO	0,915	0,923	0,832
NÚMERO CONVÊNIO	0,971	0,967	-0,397
CÓDIGO CONVENENTE	0,987	0,987	0,062
OBJETO DO CONVÊNIO	0,555	0,896	61,402
VALOR CONTRAPARTIDA	0,907	0,746	-17,737
NOME CONVENENTE	0,943	0,970	2,776
VALOR CONVÊNIO	0,817	0,841	2,858
CÓDIGO UG CONCEDENTE	0,969	0,965	-0,434
DATA FINAL VIGÊNCIA	0,857	0,889	3,776
VALOR ÚLTIMA LIBERAÇÃO	0,769	0,976	26,858
UF	0,968	0,975	0,711
NOME ÓRGÃO SUPERIOR	0,988	0,986	-0,158
DATA INÍCIO VIGÊNCIA	0,973	0,973	0,024
TIPO ENTE CONVENENTE	0,963	0,966	0,345

Os gráficos da Figura 5.27 mostram a evolução do *F1-Score*, da precisão e da sensibilidade ao longo do treinamento. É possível perceber que os três *scores*, no modelo XLM-RoBERTa, alcançaram acima de 85% rapidamente, com cerca de 400 *steps*. Por outro lado, no modelo BERTimbau, apenas a sensibilidade alcançou o mesmo patamar em um número de *steps* semelhantes. O *F1-Score* e a precisão alcançaram apenas em 3400 *steps*, o que evidencia as melhorias no pré-treinamento do modelo BERT implementadas pelo modelo RoBERTa, visto que os dois modelos foram treinados com o mesmo conjunto de dados.

Os gráficos da Figura 5.28 ilustram a evolução das perdas para os dois modelos ao longo do treinamento. É possível ver que, após o pico inicial, as perdas do *Transformer* no modelo BERTimbau oscilam ao longo do treinamento, enquanto as perdas do REN decaem devagar. Já no modelo XLM-RoBERTa, tanto as perdas do *Transformer* quanto as perdas do REN decaem rapidamente após o pico inicial. Por meio desses gráficos, é possível comprovar os aprimoramentos no pré-treinamento do modelo BERT propostos por Liu et al. (2019)

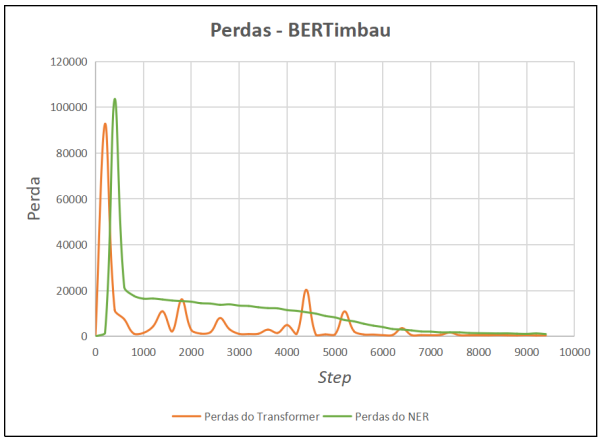


(a) Modelo BERTimbau.

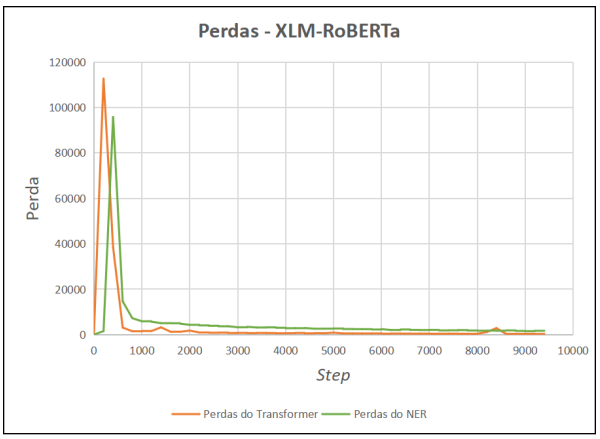


(b) Modelo XLM-RoBERTa.

Figura 5.27 – Evolução dos *scores* durante o treinamento.



(a) Modelo BERTimbau.



(b) Modelo XLM-RoBERTa.

Figura 5.28 – Evolução das perdas durante o treinamento

Tabela 5.11 – Resultados do modeo BERTimbau.

Desempenho Geral			**Desempenho por Entidade**				**Perdas do *Transformer***	**Perdas do REN**
F1-Score	***Precision***	***Recall***	**Entidade**	***F1-Score***	***Precision***	***Recall***		
0,940	0,944	0,935	NÚMERO CONVÊNIO	0,971	0,979	0,975	498,729	1370,087
			NOME ÓRGÃO SUPERIOR	0,988	0,992	0,990		
			CÓDIGO UG CONCEDENTE	0,969	0,976	0,972		
			NOME CONVENENTE	0,943	0,955	0,949		
			CÓDIGO CONVENENTE	0,987	0,991	0,989		
			VALOR CONVÊNIO	0,817	0,841	0,829		
			DATA INÍCIO VIGÊNCIA	0,973	0,984	0,979		
			NÚMERO PROCESSO DO CONVÊNIO	0,969	0,980	0,974		
			UF	0,968	0,880	0,922		
			VALOR CONTRAPARTIDA	0,907	0,899	0,903		
			DATA FINAL VIGÊNCIA	0,857	0,769	0,810		
			OBJETO DO CONVÊNIO	0,555	0,429	0,484		
			NOME ÓRGÃO CONCEDENTE	0,978	0,979	0,979		
			VALOR ÚLTIMA LIBERAÇÃO	0,769	0,786	0,777		
			TIPO ENTE CONVENENTE	0,963	0,980	0,972		
			NOME MUNICÍPIO	0,915	0,780	0,842		
			NÚMERO ORIGINAL	0,723	0,750	0,736		
			CÓDIGO ÓRGÃO CONCEDENTE	0,962	0,974	0,968		
			NOME UG CONCEDENTE	0,882	0,938	0,909		
			TIPO INSTRUMENTO	0,929	1,000	0,963		
			DATA ÚLTIMA LIBERAÇÃO	0,600	0,214	0,316		
			CÓDIGO SIAFI MUNICÍPIO	0,000	0,000	0,000		
			SITUAÇÃO CONVÊNIO	0,000	0,000	0,000		
			CÓDIGO ÓRGÃO SUPERIOR	1,000	1,000	1,000		

Tabela 5.12 – Resultados do modelo XLM-RoBERTa.

Desempenho Geral			**Desempenho por Entidade**				**Perdas do *Transformer***	**Perdas do REN**
F1-Score	***Precision***	***Recall***	**Entidade**	***F1-Score***	***Precision***	***Recall***		
0,942	0,949	0,935	NÚMERO CONVÊNIO	0,967	0,981	0,974	369,671	1741,185
			NOME ÓRGÃO SUPERIOR	0,986	0,994	0,990		
			CÓDIGO UG CONCEDENTE	0,965	0,988	0,976		
			NOME CONVENENTE	0,970	0,952	0,961		
			CÓDIGO CONVENENTE	0,987	0,991	0,989		
			VALOR CONVÊNIO	0,841	0,827	0,834		
			DATA INÍCIO VIGÊNCIA	0,973	0,983	0,978		
			NÚMERO PROCESSO DO CONVÊNIO	0,970	0,977	0,974		
			UF	0,975	0,880	0,925		
			VALOR CONTRAPARTIDA	0,746	0,818	0,780		
			DATA FINAL VIGÊNCIA	0,889	0,925	0,906		
			OBJETO DO CONVÊNIO	0,896	0,722	0,800		
			NOME ÓRGÃO CONCEDENTE	0,560	0,456	0,503		
			VALOR ÚLTIMA LIBERAÇÃO	0,976	0,981	0,978		
			TIPO ENTE CONVENENTE	0,966	0,987	0,977		
			NOME MUNICÍPIO	0,923	0,792	0,853		
			NÚMERO ORIGINAL	0,752	0,730	0,741		
			CÓDIGO ÓRGÃO CONCEDENTE	0,949	0,974	0,962		
			NOME UG CONCEDENTE	0,941	1,000	0,970		
			TIPO INSTRUMENTO	0,545	0,214	0,308		
			DATA ÚLTIMA LIBERAÇÃO	0,857	0,923	0,889		
			CÓDIGO SIAFI MUNICÍPIO	0,000	0,000	0,000		
			SITUAÇÃO CONVÊNIO	0,000	0,000	0,000		
			CÓDIGO ÓRGÃO SUPERIOR	1,000	0,667	0,800		

6 Conclusões

O objetivo deste trabalho foi treinar diferentes modelos de *Transformers* para serem capazes de reconhecer entidades nomeadas em publicações do DOU. Para alcançar esse objetivo foi necessária a construção de uma base de publicações anotadas de forma a possibilitar o aprendizado dos modelos. Com o propósito de tornar a anotação dos dados eficiente e eficaz e, consequentemente, prover um conjunto de dados com qualidade para o aprendizado dos modelos, a construção da base de publicações anotadas foi estruturada em quatro etapas.

Na primeira etapa, viu-se a importância de se conhecer as entidades a serem reconhecidas nos textos das publicações. Através do mapeamento das entidades, foi possível decidir quais eram mais adequadas para utilização nas etapas de busca e validação das publicações, assim como verificar as diferentes formas com que cada entidade pode aparecer nos textos. Foi constatado que as entidades que possuem unicidade, ou seja, que costumam ser bons identificadores de um convênio, são as mais eficientes na busca por publicações. Por outro lado, concluiu-se que as entidades que aparecem com maior frequência nas publicações são as melhores para validação das mesmas.

Na segunda e terceira etapas, tratou-se da técnica utilizada na busca de publicações e na validação das mesmas. Viu-se que essas etapas, em conjunto com a etapa de anotação, produziram uma base com 192.900 publicações anotadas. Entre essas publicações, foi observada uma variação significativa na quantidade de anotações para cada entidade. Enquanto a entidade Data da Publicação teve apenas uma anotação, a entidade Tipo Ente Convenente teve mais de 2 milhões de anotações.

Apesar do desbalanceamento das entidades, foi verificado um desempenho muito bom após o treinamento dos modelos. Apesar de terem obtido um F1-Score muito similar, com o modelo BERTimbau alcançando 0,94 e o XLM-RoBERTa 0,949, foi observado um contraste quando verificado o desempenho por entidade. Viu-se que para três entidades o modelo XLM-RoBERTa teve um desempenho significativamente melhor comparado ao BERTimbau. Por outro lado, em outras três entidades o modelo BERTimbau se sobressaiu. Outro ponto relevante foi o melhor desempenho do modelo XLM-RoBERTa em entidades mais complexas, como o Objeto do Convênio. Porém, para a entidade Nome Órgão Concedente, que é crítica para investigações de fraudes, o modelo BERTimbau obteve um resultado significativamente melhor comparado ao modelo XLM-RoBERTa.

Para trabalhos futuros pretende-se realizar o balanceamento da base de publicações anotadas a fim de melhorar o aprendizado do modelo. Outro possível trabalho futuro é estudar a influência de cada entidade no desempenho do modelo. Para alcançar isso deseja-

se gerar subconjuntos de dados a partir do conjunto de dados construído neste trabalho, onde a cada subconjunto gerado, incrementaria-se uma entidade para verificar a variação de desempenho do modelo.

Apesar deste trabalho ter sido realizado no escopo de publicações de convênios, ele é um ponto de partida para outros trabalhos que irão abranger mais instrumentos públicos, como licitações e contratos, e também outros tipos de entidades relevantes, como servidores públicos. Além disso, pretende-se incluir no modelo a capacidade de compreender o vínculo entre os convênios, as licitações e os contratos.

Referências

ALBANAZ, J. O. L. **Reconhecimento de Entidades Nomeadas em resultados de licitações publicados em Diários Oficiais**. 2020. F. 6. Especialização – Universidade Federal do Paraná, Curitiba. Citado nas pp. 13, 36, 38.

ALLES, V. J.; GIOZZA, W. F.; OLIVEIRA ALBURQUERQUE, R. de. Natural language processing to classify named entities of the Brazilian Union Official Diary. In: 2018 13th Iberian Conference on Information Systems and Technologies (CISTI). 2018. P. 1–6. DOI: 10.23919/CISTI.2018.8399215. Citado nas pp. 36, 38.

ALVAREZ, R. **Kaldi now offers TensorFlow integration**. 2017. Disponível em: <https://developers.googleblog.com/2017/08/kaldi-now-offers-tensorflow-integration.html>. Acesso em: 25 jan. 2023. Citado na p. 16.

BA, J. L.; KIROS, J. R.; HINTON, G. E. Layer Normalization, 2016. arXiv: 1607.06450 [stat.ML]. Citado na p. 32.

BRASIL. **Lei n. 14.133, de 1 de abril de 2021**. 2021. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14133.htm>. Acesso em: 23 jul. 2023. Citado na p. 12.

BRASIL. **Lei n. 8.666, de 21 de junho de 1993**. 1993. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/l8666cons.htm>. Acesso em: 13 abr. 2023. Citado nas pp. 12, 36.

BRASIL. **MEDIDA PROVISÓRIA Nº 961, DE 6 DE MAIO DE 2020**. 2020. Disponível em: <https://www.in.gov.br/web/dou/-/medida-provisoria-n-961-de-6-de-maio-de-2020-255615815>. Acesso em: 18 abr. 2023. Citado na p. 12.

CGU. **CGU monitora aplicação dos recursos federais repassados a estados e municípios**. 2023. Disponível em: <https://www.gov.br/cgu/pt-br/coronavirus/cgu-monitora-aplicacao-dos-recursos-federais-repassados-a-estados-e-municipios>. Acesso em: 12 abr. 2023. Citado na p. 12.

CONNEAU, A.; KHANDELWAL, K.; GOYAL, N.; CHAUDHARY, V.; WENZEK, G.; GUZMÁN, F.; GRAVE, E.; OTT, M.; ZETTLEMOYER, L.; STOYANOV, V. Unsupervised Cross-lingual Representation Learning at Scale. In: PROCEEDINGS of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, jul. 2020. P. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. Disponível em: <https://aclanthology.org/2020.acl-main.747>. Citado na p. 51.

COSTA, R. P. da. **Reconhecimento de Entidades Nomeadas em Textos Informais no Domínio Legislativo**. 2023. F. 75. Mestrado – Universidade Federal de Goiás, Goiânia. Citado nas pp. 37–39.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: PROCEEDINGS of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, jun. 2019. P. 4171–4186. DOI: 10.18653/v1/N19-1423. Disponível em: <https://aclanthology.org/N19-1423>. Citado na p. 49.

FERRI, J. **Abordagem modular baseada em dicionário para reconhecimento de entidades nomeadas através de associação aproximada**. 2016. F. 72. Mestrado – Universidade Federal do Paraná, Curitiba. Citado nas pp. 37, 39.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. P. 770–778. DOI: 10.1109/CVPR.2016.90. Citado na p. 32.

HONNIBAL, M.; MONTANI, I.; VAN LANDEGHEM, S.; BOYD, A. spaCy: Industrial-strength Natural Language Processing in Python, 2020. DOI: 10.5281/zenodo.1212303. Citado na p. 48.

HUGGING FACE. **RoBERTa**. 2023. Disponível em: <https://huggingface.co/docs/transformers/model_doc/roberta>. Acesso em: 17 jun. 2023. Citado na p. 50.

IBM. **What is natural language processing (NLP)?** Disponível em: <https://www.ibm.com/topics/natural-language-processing>. Acesso em: 23 jan. 2023. Citado nas pp. 15, 16, 18.

IMPRENSA NACIONAL. **Diário Oficial da União**: Pesquisa. Disponível em: <https://www.in.gov.br/consulta/-/buscar/>. Acesso em: 29 mai. 2023. Citado nas pp. 41, 42, 44.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. Third Edition draft, 2023. Acesso em: 30 jan. 2023. Citado nas pp. 15, 17–22, 24–35, 50.

KARATAS, G. **Speech Recognition: Everything You Need to Know in 2023**. 2023. Disponível em: <https://research.aimultiple.com/speech-recognition/>. Acesso em: 17 jun. 2023. Citado na p. 15.

LIMA, M.; SILVA, R.; MENDES, F.; CARVALHO, L.; ARAUJO, A.; VIDAL, F. Inferring about fraudulent collusion risk on Brazilian public works contracts in official texts using a Bi-LSTM approach. In: p. 1580–1588. DOI: 10.18653/v1/2020.findings-emnlp.143. Citado na p. 13.

LIMA, M. C. **Deep Vacuity: Detecção e Classificação Automática de Padrões com Risco de Conluio em Dados Públicos de Licitações de Obras**. 2021. Diss. (Mestrado). DOI: https://repositorio.unb.br/handle/10482/42026. Citado nas pp. 12, 13.

LIU, Y.; OTT, M.; GOYAL, N.; DU, J.; JOSHI, M.; CHEN, D.; LEVY, O.; LEWIS, M.; ZETTLEMOYER, L.; STOYANOV, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019. cite arxiv:1907.11692. Disponível em: <http://arxiv.org/abs/1907.11692>. Citado nas pp. 50, 60.

MGISP. **Painel de Transferências**. 2023. Disponível em: <https://clusterqap2.economia.gov.br/extensions/painel-gestao-transferencias/painel-gestao-transferencias.html>. Acesso em: 12 abr. 2023. Citado na p. 12.

NAVIGLI, R. **Lecture 7: Word Sense Disambiguation**. Disponível em: <https://navigli nlp.blogspot.com/2013/05/lecture-7-word-sense-disambiguation.html>. Acesso em: 26 jan. 2023. Citado na p. 17.

NETRIN. **Web Scraping**: O que é, como funciona e para que serve? Disponível em: <https://netrin.com.br/web-scraping-o-que-e-como-funciona/>. Acesso em: 4 jun. 2023. Citado na p. 44.

OLIVEIRA LOPES, A. de. **Superfaturamento de Obras Públicas**. São Paulo: Livro Pronto, 2011. Acesso em: 12 abr. 2023. Citado na p. 12.

PORTAL DA TRANSPARÊNCIA. **Ação Orçamentária no Enfrentamento da Emergência de Saúde Pública de Importância Internacional Decorrente do Coronavírus**. 2023a. Disponível em: <https://portaldatransparencia.gov.br/programas-e-acoes/acao/21C0-enfrentamento-da-emergencia-de-saude-publica-de-importancia-internacional-decorrente-do-coronavirus?ano=2020>. Acesso em: 12 abr. 2023. Citado na p. 12.

PORTAL DA TRANSPARÊNCIA. **Convênios**. Disponível em: <https://portaldatransparencia.gov.br/download-de-dados/convenios>. Acesso em: 29 mai. 2023. Citado nas pp. 40, 44.

PORTAL DA TRANSPARÊNCIA. **Dicionário de Dados - Convênios**. Disponível em: <https://portaldatransparencia.gov.br/pagina-interna/603415-dicionario-de-dados-convenios>. Acesso em: 29 mai. 2023. Citado na p. 40.

PRAKASH, A. **What is transformer architecture and how does it power ChatGPT?** 2023. Disponível em: <https://www.thoughtspot.com/data-trends/ai/what-is-transformer-architecture-chatgpt>. Acesso em: 25 jun. 2023. Citado na p. 13.

PUSTEJOVSKY, J.; STUBBS, A. **Natural Language Annotation for Machine Learning**. O'Reilly Media, Incorporated, 2012. (A Guide to corpus-building for applications, v. 9,p. 878). ISBN 9781449306663. Disponível em: <https://books.google.com.br/books?id=QtzmqamXxx4C>. Citado nas pp. 15, 48.

RAMESH, A. **Using Natural Language Processing to Power Chatbots**. 2019. Disponível em: <https://discover.bot/bot-talk/behind-the-scenes-using-nlp-to-power-chatbot/>. Acesso em: 30 jan. 2023. Citado na p. 18.

SALAS, J.; BARROS VIDAL, F. de; MARTINEZ-TRINIDAD, F. Deep Learning: Current State. **IEEE Latin America Transactions**, v. 17, n. 12, p. 1925–1945, 2019. DOI: 10.1109/TLA.2019.9011537. Citado na p. 13.

SANTANA, J. B. de. **Desenvolvimento e Análise de Corpus para Reconhecimento de Entidades Nomeadas em Relatórios de Inteligência Financeira**. 2020. F. 104. Mestrado – Universidade Federal de Santa Catarina, Florianópolis. Citado nas pp. 37, 39.

SANTOS NAKAMURA, A. L. dos. A infraestrutura e a corrupção no Brasil. **Revista Brasileira de Estudos Políticos**, n. 117, p. 97–126, jul. 2018. Citado na p. 12.

SILVA RODRÍGUEZ, M. M. M. da; BEZERRA, B. L. D. Processamento de Linguagem Natural para Reconhecimento de Entidades Nomeadas em Textos Jurídicos de Atos Administrativos (Portarias). In: 1. REVISTA de Engenharia e Pesquisa Aplicada: Edição Especial em Ciência de Dados e Analytics. 2020. v. 5, p. 67–77. DOI: 1025286/repa.v5i1.1204. Citado nas pp. 36, 38.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: 9TH Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear). 2020. Citado na p. 50.

SOUZA, K. **Os 4 tipos de fraudes mais comuns nas licitações e contratos do Covid-19**. 2023. Disponível em: <https://3rcapacita.com.br/artigo/principais-tipos-de-fraudes-constatadas-nas-licitacoes-e-contratos-do-covid-19>. Acesso em: 12 abr. 2023. Citado na p. 12.

SPACY. **displaCy**. Disponível em: <https://spacy.io/usage/visualizers#ent>. Acesso em: 17 jun. 2023. Citado na p. 48.

SPACY. **Linguistic Features**. Disponível em: <https://spacy.io/usage/linguistic-features#pos-tagging>. Acesso em: 25 jan. 2023. Citado nas pp. 16, 18.

SPACY. **PhraseMatcher**. Disponível em: <https://spacy.io/api/phrasematcher>. Acesso em: 17 jun. 2023. Citado na p. 48.

TUCKER, A. B. **Overview of NLP: Issues and Strategies**. 2002. Disponível em: <https://tildesites.bowdoin.edu/~allen/nlp/nlp1.html>. Acesso em: 2 mar. 2023. Citado na p. 23.

VAN RIJSBERGEN, C. **Information Retrieval**. Butterworths, 1975. ISBN 9780408707176. Disponível em: <https://books.google.com.br/books?id=EJ2PQgAACAAJ>. Citado na p. 35.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. Attention Is All You Need. eng, 2017. Citado nas pp. 27, 34.

VOXCO. **Sentiment Analysis helps improve Customer Experience**. Disponível em: <https://www.voxco.com/blog/sentiment-analysis-helps-improve-customer-experience/>. Acesso em: 24 jan. 2023. Citado na p. 17.

WOLF, T.; DEBUT, L.; SANH, V.; CHAUMOND, J.; DELANGUE, C.; MOI, A.; CISTAC, P.; RAULT, T.; LOUF, R.; FUNTOWICZ, M.; DAVISON, J.; SHLEIFER, S.; PLATEN, P. von; MA, C.; JERNITE, Y.; PLU, J.; XU, C.; LE SCAO, T.; GUGGER, S.; DRAME, M.; LHOEST, Q.; RUSH, A. Transformers: State-of-the-Art Natural Language Processing. In: PROCEEDINGS of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Online: Association for Computational Linguistics, out. 2020. P. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. Disponível em: <https://aclanthology.org/2020.emnlp-demos.6>. Citado nas pp. 13, 49.

ZHU, X. J. **CS 838-1: Advanced Natural Language Processing**. 2007. Disponível em: <https://pages.cs.wisc.edu/~jerryzhu/cs838/cs838.html>. Acesso em: 8 mar. 2023. Citado nas pp. 25, 26.