Universidade de Brasília
Faculdade de Tecnologia

Reconhecimento de Entidades Nomeadas baseado em Transformers aplicado no relacionamento textual de licitações públicas e publicações em diários oficiais

Tiago Ferreira Candido

PROJETO FINAL DE CURSO

ENGENHARIA DE CONTROLE E AUTOMAÇÃO

Brasília

2023

Universidade de Brasília
Faculdade de Tecnologia

Reconhecimento de Entidades Nomeadas baseado em Transformers aplicado no relacionamento textual de licitações públicas e publicações em diários oficiais

Tiago Ferreira Candido

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

Candido, Tiago Ferreira.
C217r Reconhecimento de Entidades Nomeadas baseado em Transformers aplicado no relacionamento textual de licitações públicas e publicações em diários oficiais / Tiago Ferreira Candido; orientador Flávio de Barros Vidal. -- Brasília, 2023.
68 p.

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

1. Processamento de linguagem natural. 2. Transformers. 3. Dados públicos. 4. BERT. I. Vidal, Flávio de Barros, orient. II. Título

Universidade de Brasília
Faculdade de Tecnologia

Reconhecimento de Entidades Nomeadas baseado em Transformers aplicado no relacionamento textual de licitações públicas e publicações em diários oficiais

Tiago Ferreira Candido

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

Trabalho aprovado. Brasília, 21 de Dezembro de 2023:

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

Prof. Dr. Tiago Alves da Fonseca,
FGA/UnB
Examinador interno

Prof. Dr. Marcus Vinícius Chaffim Costa,
FGA/UnB
Examinador interno

Brasília
2023

Dedico este trabalho primeiramente a minha família. Agradeço a meus pais e a minha irmãzinha que, com todo carinho e dedicação, me proporcionaram todas as ferramentas para que este sonho se tornasse possível. Em segundo lugar, dedico a todos os docentes que fizeram parte da minha jornada até a graduação.

Agradecimentos

Agradeço a Deus por sempre tomar conta de mim, me proporcionar saúde e uma família maravilhosa. A meus pais pela educação, amor e todo o esforço que resultou na formação do meu caráter, vocês foram essenciais para minha graduação, para todas as conquistas que vieram e estão por vir. A todos os colegas que tive, vocês foram responsáveis por tornar o ambiente universitário um lugar mais confortável e divertido. Por fim agradeço a todo o corpo de profissionais que compõem essa incrível universidade, foi um prazer estudar na Universidade de Brasília.

Resumo

O avanço no campo do Processamento de Linguagem Natural (PLN) tem proporcionado resultados significativos no reconhecimento de elementos textuais, desempenhando papel crucial em diversas atividades relacionadas à identificação e tratamento de conteúdo textual. Este trabalho de graduação aborda a necessidade de aprimorar as técnicas empregadas nesse processo, embora o reconhecimento de elementos textuais por meio do PLN tenha evoluído, a ausência de técnicas mais elaboradas e complexas é notória, sendo atribuída, em grande parte, às restrições de poder computacional. Devido à complexidade crescente das estruturas linguísticas e a vastidão dos dados textuais exigem abordagens mais avançadas para garantir a precisão e eficiência no processamento. No cenário brasileiro onde vastas quantidades de informações públicas circulam de maneira granular e desorganizada pelos três poderes, torna-se essencial a criação ou adaptação de técnicas para coletar e tratar informações relevantes para diversos setores da sociedade. Nesse contexto, destaca-se que a necessidade de aprimorar a capacidade de identificação e classificação de informações em grandes conjuntos de dados textuais é premente, especialmente em ambientes governamentais onde a transparência é crucial. Diante desse contexto, este trabalho propõe a realização de estudos e desenvolvimentos utilizando técnicas de alto desempenho e custo computacional baseadas em modelos *Transformers* para a realização do reconhecimento de entidades nomeadas por relação textual em licitações públicas e publicações em diários oficiais. A ênfase está na melhoria do processo de classificação de elementos textuais em informações públicas, com foco específico em licitações brasileiras. Dessa forma, este trabalho busca contribuir para a evolução das técnicas de reconhecimento de elementos textuais, proporcionando avanços significativos no tratamento de informações públicas, com potenciais benefícios para diversos setores da sociedade, especialmente no contexto das licitações brasileiras.

Palavras-chave: Processamento de linguagem natural. Transformers. Dados públicos. BERT.

Abstract

Advances in Natural Language Processing (NLP) have provided significant results in recognizing textual elements, playing a crucial role in various activities related to identifying and processing textual content. This undergraduate work addresses the need to improve the techniques used in this process. Although the recognition of textual elements through PLN has evolved, the absence of more elaborate and complex techniques is notable, being attributed, in large part, to power restrictions. Due to the increasing complexity of linguistic structures and the vastness of textual data, more advanced approaches are required to ensure accuracy and efficiency in processing. In the Brazilian scenario, where vast amounts of public information circulate in a granular and disorganized manner across the three powers, it is essential to create or adapt techniques to collect and process relevant information for different sectors of society. In this context, it is highlighted that the need to improve the ability to identify and classify information in large sets of textual data is pressing, especially in government environments where transparency is crucial. Given this context, this work proposes carrying out studies and developments using high-performance and computationally expensive techniques based on *Transformers* models to recognize entities named by textual relationship in public tenders and publications in official gazettes. The emphasis is on improving the process of classifying textual elements in public information, specifically Brazilian tenders. Due to the vast amount of data available in this context, more advanced approaches are required, aiming at efficiency and effectiveness in processing this information. In this way, this work seeks to contribute to the evolution of techniques for recognizing textual elements, providing significant advances in processing public information, with potential benefits for different sectors of society, especially in the context of Brazilian tenders.

Keywords: Natural language processing. Transformers. Public data. BERT.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

1.1 Motivações e Justificativas

A corrupção no Brasil tem histórico de ser um desafio persistente. Uma das práticas mais recorrentes é a fraude em licitações e contratos públicos, conforme Júnior, Filho e Cabral (2023). Esse problema vem corroendo a integridade do setor público, minando a eficiência das políticas governamentais e prejudicando a alocação adequada dos recursos públicos. Diversos estudos acadêmicos apontam para a gravidade desse cenário. Segundo Fortini e Motta (2016), a corrupção em licitações é um dos principais obstáculos para o desenvolvimento socioeconômico de um país, pois atinge de forma direta toda a população, e de maneira ainda mais brutal sua camada economicamente mais frágil, devido ao fato de que os recursos públicos não são destinados a suprir suas carências. A quantidade elevada de dinheiro envolvido, a dificuldade em detectar fraudes de forma eficiente e a falta de uma legislação rigorosa para punir os responsáveis também contribui para a perpetuação desse problema como é abordado por Nakamura (2018).

Com o avanço da pandemia de Covid-19 em 2020, no Brasil, houve a flexibilização da legislação vigente de licitações e contratos, por meio de medida provisória, conforme disposto em Brasil (2020). Embora necessária em um momento emergencial, esta decisão aumentou os riscos de corrupção e favoreceu práticas fraudulentas em processos licitatórios. A urgência em adquirir equipamentos médicos, insumos hospitalares e serviços de saúde trouxe consigo a pressão por rapidez nas contratações, o que acabou comprometendo a devida diligência na análise das empresas contratadas, segundo Brito e Costa (2021). Conforme o Portal da Transparência (2023), somente em 2020 o governo federal dispôs de mais de R$ 40 bilhões no combate da Covid-19. Segundo a CGU (2023), em um período de dois anos de atuação direta, de abril de 2020 a abril de 2022, analisando um montante de contratações e licitações de R$ 5,87 bilhões, o prejuízo potencial decorrente do desdobramento das investigações é de pouco mais de R$ 300 milhões.

No contexto global de investigação de fraudes e conluio em licitações e contratos públicos, abordagens metodológicas abrangentes têm sido desenvolvidas para enfrentar essa questão complexa. Smith e Johnson (2018) exploraram a detecção de conluio em leilões de aquisição pública usando técnicas de aprendizado de máquina. Chen e Liu (2019) discutiram a detecção de anomalias em dados de compras públicas por meio de métodos estatísticos. Min Zhang e Zhou (2020) apresentaram uma abordagem de análise de redes para a detecção de fraudes em licitações públicas.

No âmbito nacional, o processo de investigação de fraudes em licitações e contratos

públicos se apoia em conhecimentos adquiridos ao longo dos anos. Enquanto certos tipos de fraudes possuem metodologias de investigação bem estabelecidas, como o superfaturamento de obras públicas, outros, como o conluio e a formação de cartel, exigem abordagens manuais e experientes, conforme enfatizado por Marcos Cavalcanti Lima (2021). Para casos de licitações, a Lei 8.666 em BRASIL (1993) e a Lei 14.133 em BRASIL (2021) exigem que atos oficiais sejam publicados no Diário Oficial da União (DOU). Essa fonte de dados, embora rica em informações, é desafiadora devido à falta de padronização e à alta quantidade de publicações por licitação. A aplicação de técnicas de aprendizado de máquina nas publicações do DOU é uma área pouco explorada. Em 2020, M. Lima et al. (2020) desenvolveram um modelo de classificação de publicações do DOU como parte da ferramenta Deep Vacuity, visando detectar fraudes e conluios em licitações de obras públicas no Brasil. Como destacado por Marcos Cavalcanti Lima (2021), essa ferramenta visa capacitar investigadores da Polícia Federal a identificar fraudes em licitações públicas por meio de técnicas de aprendizado de máquina.

Atualmente, a identificação de componentes nos textos das publicações no DOU ocorre por meio de uma abordagem de correspondência direta usando expressões regulares. No entanto, conforme mencionado anteriormente, as publicações no DOU carecem de um formato de escrita uniforme, o que limita a eficácia das expressões regulares como método.

Além da falta de uniformidade nas publicações, também se deparam com complexidades inerentes ao processamento e interpretação da linguagem natural, conforme destacado por Albanaz (2020). Questões como a ambiguidade de palavras (polissemia), diferentes palavras com significados semelhantes (sinonímia) e palavras pouco comuns (raridade) são nuances linguísticas que demandam uma análise mais avançada. No campo do Processamento de Linguagem Natural (PLN), a arquitetura Transformer surgiu como uma inovação notável, de acordo com as observações de Wolf et al. (2020). Essa arquitetura superou modelos anteriores, como Redes Neurais Recorrentes (RNNs) e Redes Neurais Convolucionais (CNNs), conforme também apontado por Salas, Barros Vidal e Martinez-Trinidad (2019), solidificando sua presença graças ao sucesso de modelos como BERT e GPT, sendo este último o núcleo do ChatGPT, conforme indicado por Prakash (2023).

O Transformer traz benefícios marcantes em comparação com outras abordagens. Sua eficiência e agilidade no treinamento são notáveis devido à viabilidade de treinamento em paralelo. A habilidade de capturar contextos extensos também é uma característica essencial, como delineado por Wolf et al. (2020). Além disso, o Transformer permite o pré-treinamento de modelos em um vasto corpo de texto genérico, seguido pelo refinamento do treinamento para tarefas específicas.

1.2 Objetivos

1.2.1 Objetivo Geral

O trabalho tem como objetivo principal utilizar técnicas processamento de linguagem natural para classificar publicações de naturezas distintas, advindas da API de Compras Governamentais (2023) e do DOU em Imprensa Nacional (2023), fazendo uso de modelos de Transformers pré-treinados. A finalidade é aprimorar a transparência e eficiência dos procedimentos no contexto das contratações públicas.

1.2.2 Objetivos Específicos

Diante das complexidades inerentes aos processos licitatórios, é imperativo aprimorar as ferramentas e técnicas disponíveis para a detecção e prevenção de fraudes, conluio e outras práticas irregulares que possam comprometer a integridade desses procedimentos. Dentro desse cenário, os objetivos específicos deste estudo são os seguintes:

- **Criação de uma base de dados integrada e rotulada:** Formação de um conjunto de dados unificado, relativos a processos licitatórios, composto por publicações do DOU e da API de Compras Governamentais (2023). Esse conjunto de dados, após agrupado e rotulado, servirá como base para o desenvolvimento e treinamento dos modelos de Transformers.
- **Treinamento dos modelos de Transformers:** Utilizar técnicas de aprendizado de máquina supervisionado, por meio de modelos de transformers, para classificar os dados provenientes do site compras com informações presentes no DOU. Essa etapa visa identificar padrões, relações e correspondências entre publicações, simplificando a associação desses elementos.
- **Contribuição para o combate a fraudes:** Ao estabelecer uma associação precisa entre publicações de processos licitatórios de naturezas distintas, busca-se fortalecer as medidas de controle e monitoramento, proporcionando uma ferramenta eficaz para a identificação de práticas fraudulentas e conluio nos processos licitatórios do governo brasileiro.

Ao finalizar esses objetivos, espera-se que este estudo possa oferecer uma abordagem robusta para o aprimoramento da transparência e da integridade das contratações públicas. Além disso, pretende-se contribuir para o desenvolvimento de uma base de conhecimento que possa ser explorada na área de análise de dados governamentais e detecção de irregularidades em processos licitatórios.

1.3 Organização do Trabalho

O trabalho está organizado em diversas seções que visam proporcionar uma compreensão abrangente e estruturada do estudo realizado. A seguir, é descrito o conteúdo de cada seção:

- **Capítulo 2 - Fundamentos Teóricos:** Nesta seção, serão apresentados os conceitos e conhecimentos essenciais para a compreensão dos temas abordados neste trabalho. Este segmento explora os conceitos fundamentais que abrangem aprendizado de máquina, processamento de linguagem natural e a transferência de conhecimento para modelos Transformer pré-treinados, chegando às métricas utilizadas para avaliar o desempenho na tarefa de classificação.
- **Capítulo 3 - Trabalhos Relacionados:** Nesta seção, serão examinadas pesquisas e estudos pertinentes que tratam da classificação de dados públicos, empregando técnicas avançadas de PLN em contextos semelhantes. Serão analisadas as abordagens metodológicas, as técnicas empregadas e os resultados alcançados por outros pesquisadores. Essa revisão visa posicionar nosso estudo no contexto já estabelecido, identificando lacunas ou oportunidades para contribuições originais.
- **Capítulo 4 - Metodologia:** Este capítulo detalha a metodologia adotada para alcançar os objetivos propostos. Serão descritas as etapas para a coleta e preparação dos dados provenientes do DOU e do site de compras governamentais. Além disso, será explicado sobre os modelos de Transformers que serão empregados para classificar as publicações, facilitando a associação entre publicações de processos licitatórios de fontes diferentes.
- **Capítulo 5 - Resultados:** Nesta seção, serão apresentados os resultados obtidos desde a criação da base de dados rotulada até a aplicação dos modelos de Transformers para o treinamento dos modelos. Serão discutidas a formação das classes, os padrões identificados e as correspondências entre as publicações. Também serão apresentadas métricas de avaliação para verificar a eficácia das abordagens propostas.
- **Capítulo 6 - Conclusões:** No último capítulo, é feito uma síntese dos principais achados deste estudo. Serão discutidos os resultados associados a implementação da base de dados, e sua posterior utilização para o treinamento dos modelos de Transformers, além das possibilidades de contribuição para a transparência em processos licitatórios, destacando as implicações práticas e as limitações encontradas. Ainda, serão delineadas possíveis direções futuras de pesquisa nesse campo.

Por fim será apresentada a seção de referências em que são listadas todas as fontes acadêmicas, literárias e eletrônicas utilizadas ao longo do trabalho. As referências seguirão as convenções de formatação apropriadas, possibilitando a rastreabilidade e a credibilidade

das informações utilizadas. Por meio dessa organização estruturada, este trabalho pretende proporcionar uma análise sobre a aplicação de técnicas avançadas de classificação de entidades textuais, baseadas em contexto, capazes de contribuir para um entendimento mais aprofundado da abordagem utilizada e fornecer percepções valiosas para a área de análise de dados públicos governamentais.

2 Fundamentos Teóricos

Aprendizado de máquina, processamento de linguagem natural e inteligência artificial são termos interligados que juntos capacitam as máquinas a compreender, aprender e interagir com o mundo humano de maneira semelhante ou até superior à capacidade humana. Neste capítulo, serão abordados os pilares fundamentais que sustentam o entendimento das técnicas empregadas neste trabalho. Compreender os conceitos essenciais de aprendizado de máquina, processamento de linguagem natural e a arquitetura dos Transformers é crucial para explorar a abordagem proposta de classificação a ser aplicada na base de publicações gerada. Na Seção 2.1, são apresentados os conceitos fundamentais sobre modelos de aprendizado de máquina e algumas técnicas associadas. Na Seção 2.2, são apresentados os conceitos sobre PLN. Na Seção 2.3, são apresentados os principais conceitos de Transformers, suas características e métodos de avaliação.

2.1 Aprendizado de Máquina

O Aprendizado de Máquina se concentra nos métodos de desenvolvimento de algoritmos e modelos capazes de aprender padrões a partir de dados e tomar decisões com base nesses padrões. Ou seja, em vez de programar regras explicitamente, os sistemas de aprendizado de máquina são alimentados com grandes quantidades de dados, permitindo que eles extraiam informações, identifiquem tendências e façam previsões ou classificações, segundo Kelleher, MacNamee e D'Arcy (2015). Os modelos de aprendizado de máquina podem ser divididos em duas categorias principais: supervisionados e não supervisionados. No aprendizado supervisionado, os modelos são treinados em dados rotulados e são utilizados para gerar modelos de classificação e regressão, enquanto no aprendizado não supervisionado, os modelos encontram padrões em dados não rotulados para realizar tarefas de agrupamento, associação ou sumarização como descrito por Faceli (2011). A Figura 2.1 ilustra as classificações de aprendizagem e suas principais funcionalidades. Para a tarefa alvo do trabalho será utilizado aprendizado de máquina supervisionado para a tarefa de classificação.

2.1.1 Aprendizado de Máquina Supervisionado

No contexto da aprendizagem supervisionada, os modelos são alimentados com dados que possuem rótulos associados, onde cada instância de dados possui um valor real de saída yi associado, visando capturar a relação entre variáveis preditoras, os dados de entrada, e a variável alvo, também chamada de rótulo como descrito por Faceli (2011). A base dos modelos supervisionados reside em expressões matemáticas que procuram aprender

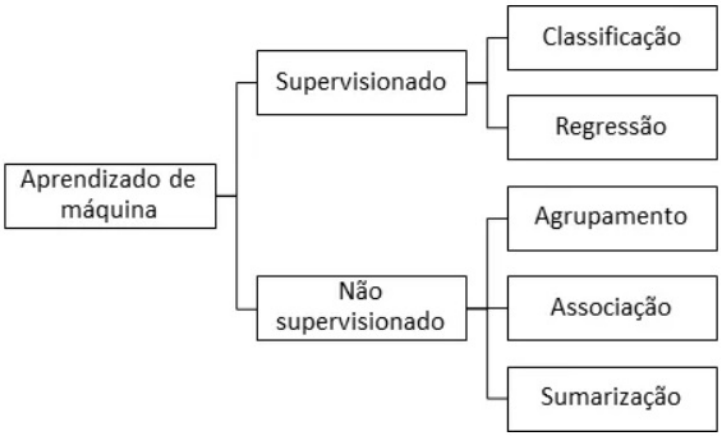

Figura 2.1 – Tipos de aprendizado de máquina.

Fonte: Candido (2023)

a função ideal $f : X \rightarrow Y$, a qual melhor traduz o problema. Como exemplo, no âmbito de classificação de documentos textuais, o conjunto X representa as entradas formadas por entidades textuais, enquanto Y representa os valores a serem previstos, ou seja, a qual classe eles pertencem.

O objetivo principal dos modelos supervisionados é aprender uma função $g : X \rightarrow Y$ com $g \approx f$, fazendo uso de dados de treinamento provenientes de um conjunto de dados. Esse conjunto de dados é geralmente dividido em três partes: treinamento, validação e teste. Os dois primeiros são utilizados para criar o modelo preditivo, e o último, para avaliar o desempenho do modelo gerado, sendo o conjunto de teste uma simulação de dados do mundo real que o modelo não encontrou durante o treinamento. O modelo se refere à função g, o qual é o resultado do processo. Existem diversas variações de modelos, que são empregados de acordo com a natureza do problema. A Figura 2.2 apresenta a ideia do aprendizado de máquina supervisionado.

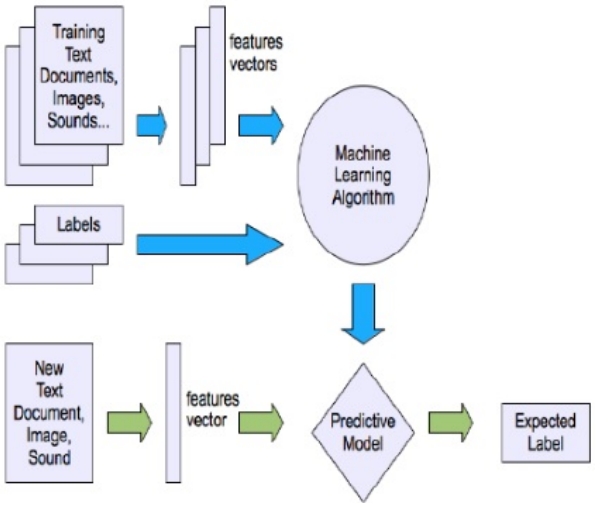

Figura 2.2 – Estrutura de aprendizado de máquina supervisionado.

Fonte: Jayanthi e Mahesh (2018)

2.1.2 Aprendizado de Máquina Não Supervisionado

Na esfera da aprendizagem não supervisionada, os modelos se deparam com dados não rotulados, nos quais identificam padrões de similaridade anteriormente desconhecidos e baseiam suas decisões na detecção da presença ou ausência de tais padrões em novos dados segundo Khanam et al. (2015). A clusterização, ou agrupamento de dados é um exemplo clássico nesse paradigma, onde o modelo encontra similaridades entre pontos de dados no conjunto e os organiza em grupos coerentes, ou seja, clusters de acordo com Benabdellah, Benghabrit e Bouhaddou (2019). As aplicações da aprendizagem não supervisionada incluem sistemas de recomendação de filmes ou músicas, detecção de anomalias e a visualização de padrões em conjuntos de dados. A Figura 2.3 apresenta a ideia do aprendizado de máquina supervisionado.

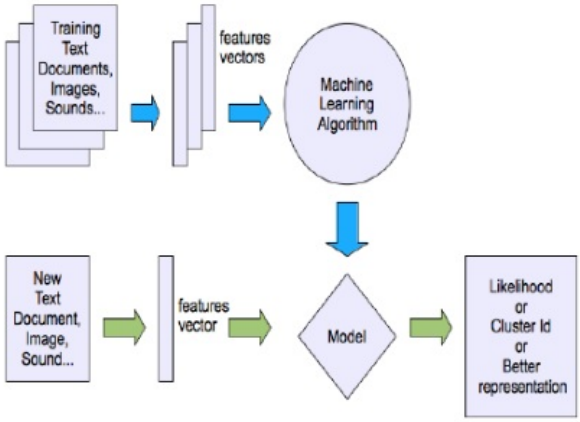

Figura 2.3 – Estrutura de aprendizado de máquina não supervisionado.

Fonte: Jayanthi e Mahesh (2018)

2.2 Processamento de Linguagem Natural

A área do Processamento de Linguagem Natural pertence ao âmbito da Inteligência Artificial e reúne abordagens oriundas tanto da Ciência da Computação quanto da Linguística, conforme demonstrado na Figura 2.4. O propósito central do PLN é explorar a interação entre computadores e a linguagem humana, empregando abordagens que variam desde aprendizado de máquina até métodos estatísticos e aprendizado profundo. Essas técnicas são aplicadas para analisar, compreender e gerar diferentes formas de expressão linguística, abrangendo texto, fala e até imagens segundo Bird, Klein e Loper (2009).

De acordo com o levantamento da IBM (2023), a criação de um software capaz de compreender plenamente a intenção pretendida em dados textuais ou de voz é uma tarefa de extrema complexidade e desafio. Isso se deve à natureza intrinsecamente ambígua da linguagem humana, bem como às suas numerosas irregularidades, como homônimos, homófonos, sarcasmo, expressões idiomáticas, metáforas, exceções gramaticais e de uso,

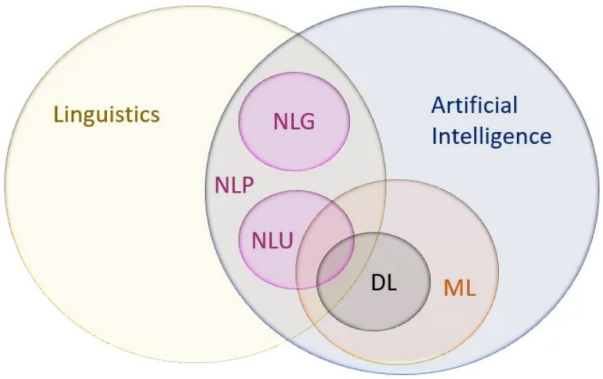

Figura 2.4 – Relação entre linguística, inteligência artificial e PLN.

Fonte: Medium (2023)

além de variações na estrutura da frase, entre outros fatores. Por essa razão, para alcançar uma compreensão computacional da linguagem humana, o campo do PLN evoluiu para abranger várias subáreas, cada uma desempenhando um papel específico na compreensão global da linguagem. A seguir, é apresentado uma breve descrição das principais subáreas do PLN:

- **O reconhecimento de fala** é a subárea que realiza a tarefa de converter, de maneira confiável, dados de voz em dados de texto, segundo IBM (2023). Os desafios do reconhecimento de fala compreendem a forma como as pessoas falam, rapidamente, juntando palavras, com ênfase e entonação variadas, em diferentes sotaques e frequentemente utilizando linguagem coloquial. Einorytė (2023) descreve o processo de reconhecimento de fala iniciando com a captação de áudio por um microfone em dispositivos como telefones e computadores. Em seguida, a tecnologia converte a gravação em informações digitais, eliminando ruídos indesejados e realizando ajustes nas características do discurso, como tom, volume e ritmo. Posteriormente, essas informações processadas são transformadas em frequências e analisadas para interpretar o discurso humano. O resultado é uma sequência de palavras correspondentes ao sinal de fala, transcrita em texto legível, a Figura 2.5 ilustra o sistema descrito.

  O processo, embora aparentemente simples, envolve complexidade, abrangendo processamento de sinais, aprendizado de máquina e processamento de linguagem natural, com a velocidade de processamento superando a capacidade humana. A precisão da saída depende da qualidade da gravação original, da complexidade do idioma e da aplicação do sistema. O reconhecimento de fala é uma tecnologia em crescimento rápido com diversas aplicações. Destacam-se o uso em sistemas de navegação veicular, assistentes virtuais como Siri e Google Assistant, transcrição médica, automação em centros de atendimento ao cliente, acessibilidade para pessoas com deficiência, tradução automática de idiomas e busca na internet por meio de comandos de voz. Esta forma de inteligência artificial automatiza processos, melhorando eficiência e precisão

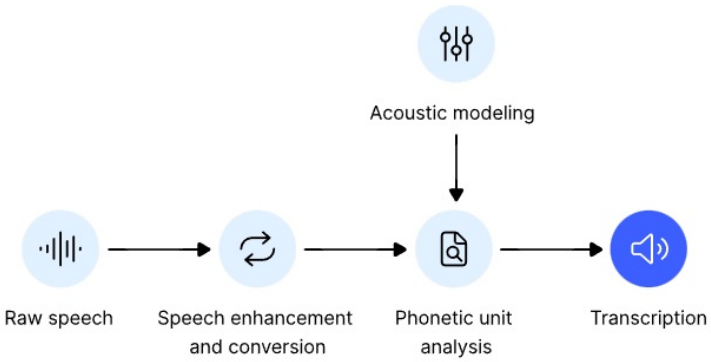

Figura 2.5 – Sistema de reconhecimento de fala.

Fonte: Einorytė (2023)

em diferentes setores, com expectativa de expansão futura.

- **Análise sintática e gramatical**, também chamada de marcação gramatical, é o procedimento de determinar a classe gramatical de uma palavra ou trecho de texto com base em seu uso e contexto. Essa é uma etapa essencial em diversas tarefas de PLN, como

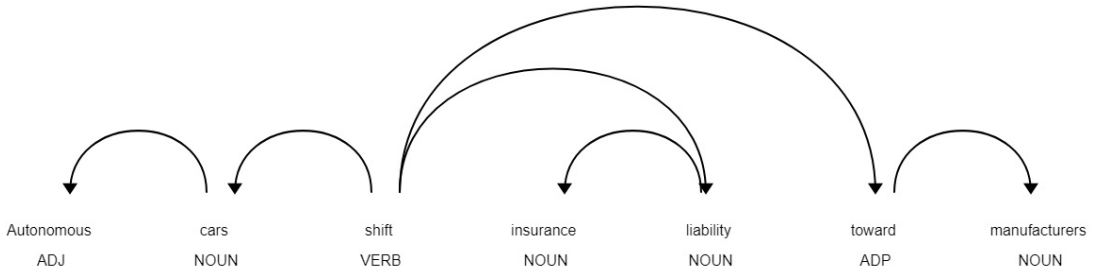

Figura 2.6 – Exemplo de análise de sintaxe e gramática.

Fonte: Spacy (2023)

análise sintática, rotulagem de função semântica e resumo de texto. De acordo com Manning, Raghavan e Schütze (2009), esse processo pode ser desmembrado em três etapas cruciais: tokenização, resolução de ambiguidade e classificação. A tokenização envolve a subdivisão de uma frase ou texto em unidades individuais, chamadas tokens. Essa fase é fundamental para permitir uma abordagem mais detalhada da linguagem, facilitando análises específicas em níveis mais baixos, como palavras isoladas. A resolução de ambiguidade surge como um desafio significativo, abordando a complexidade em que uma palavra pode apresentar várias classificações gramaticais, dependendo do contexto. A etapa final, classificação, busca atribuir uma classe gramatical a cada token. Isso muitas vezes envolve o uso de modelos de aprendizado de máquina treinados em grandes conjuntos de dados anotados. Um exemplo prático desse processo é destacado na Figura 2.6, ilustrando uma análise sintática e gramatical.

- **Análise semântica e pragmática** é a seleção do significado de uma palavra com múltiplos significados por meio de análise semântica que determina qual palavra faz mais sentido no contexto dado. Tambem conhecida como *Word Sense Disambiguation* (WSD) em PLN, é essencial em tarefas como recuperação de informações e tradução automática. O desafio reside na polissemia, onde palavras têm múltiplos significados, exigindo a compreensão do contexto para atribuir o sentido apropriado. O WSD busca resolver essa ambiguidade lexical, crucial para a interpretação precisa, segundo Agirre, Edmonds e Worden (2007) que aborda técnicas elaboradas de WSD. A Figura 2.7 ilustra um exemplo de análise semântica para extrair o sentido de uma palavra em uma entrada textual.

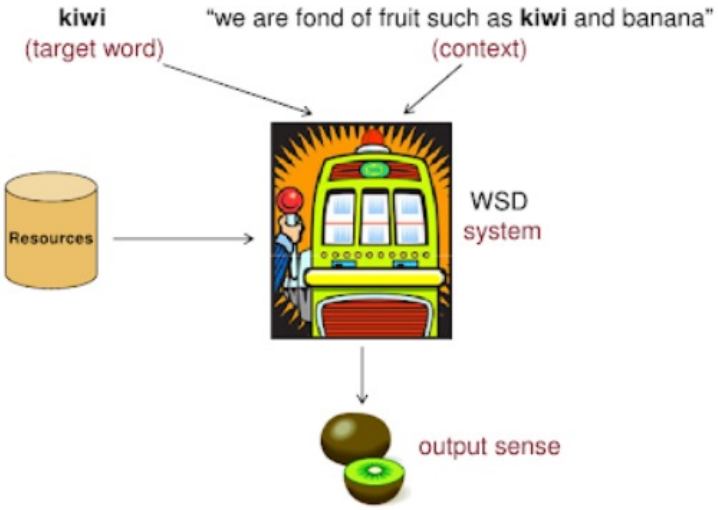

Figura 2.7 – Exemplo de analise semântica de uma palavra em um contexto específico.

Fonte: Navigli (2013)

- **O Reconhecimento de Entidades Nomeadas** (REN) visa identificar e categorizar palavras ou frases em um texto como entidades específicas, como nomes de pessoas, locais, datas, organizações e muito mais. O objetivo é extrair informações relevantes e estruturadas do texto, permitindo que os sistemas de PLN compreendam melhor o conteúdo e realizem análises mais precisas. Um exemplo básico de reconhecimento de entidades nomeadas pode ser aplicado na seguinte frase: Quando Sebastian Thrun começou a trabalhar com carros autônomos na Google em 2007, poucas pessoas fora da empresa o levaram a sério. Nessa frase, o REN identifica *Sebastian Thrun* como uma entidade do tipo pessoa, *Google* do tipo organização e *2007* do tipo data, como demonstrada na Figura 2.8. De acordo com Tjong Kim Sang e De Meulder (2003), essa categorização de entidades ajuda a organizar e extrair informações relevantes do texto, o que pode ser útil em uma ampla gama de aplicações, como resumos automáticos, análise de notícias e sistemas de busca de informações específicas.

Figura 2.8 – Exemplo do reconhecimento de entidades nomeadas em um texto.

Fonte: Spacy (2023)

- **Análise de sentimento** busca extrair qualidades subjetivas, como atitudes, emoções, sarcasmo, confusão e suspeita, do texto. Conforme descrito por Pang, Lee e Vaithyanathan (2002), essa tarefa desempenha um papel crucial em diversas aplicações, como avaliação de produtos, monitoramento de redes sociais e análise de opinião pública. Essa atividade visa determinar a polaridade emocional associada a um texto, classificando-o como positivo, negativo ou neutro, conforme exemplificado na Figura 2.9. O desafio na análise de sentimento reside na compreensão das nuances da lin-

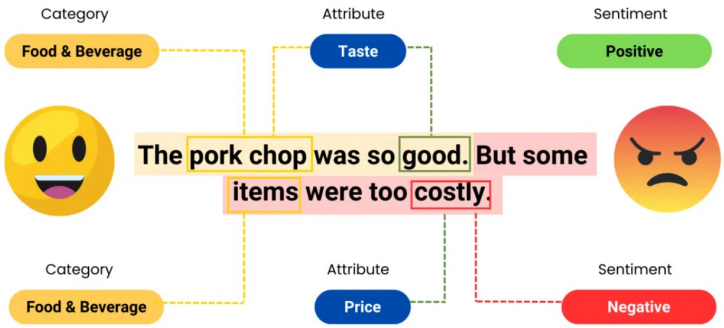

Figura 2.9 – Exemplo de análise de sentimento.

Fonte: Thirdeye (2023)

guagem, considerando expressões idiomáticas, ironia e contexto cultural. Métodos de aprendizado de máquina são frequentemente empregados para treinar modelos capazes de discernir o tom emocional de uma declaração.

- **Geração de linguagem natural**, NLG do inglês, é uma área essencial em PLN que visa criar automaticamente textos sem intervenção direta. Ela é por vezes vista como o oposto do reconhecimento de fala ou transcrição de voz para texto por ter o objetivo de transformar informações estruturadas em linguagem humana. A Figura 2.10 ilustra o processo de NLG em um *chatbot*. Suas principais aplicações são em assistentes virtuais, resumos automáticos e relatórios automatizados, a NLG enfrenta o desafio de produzir textos fluentes e semanticamente corretos, segundo Reiter e Dale (2000). Modelos avançados, como os baseados em linguagem, são comumente empregados para gerar frases e parágrafos completos de forma genérica, coerentes e em qualquer língua.

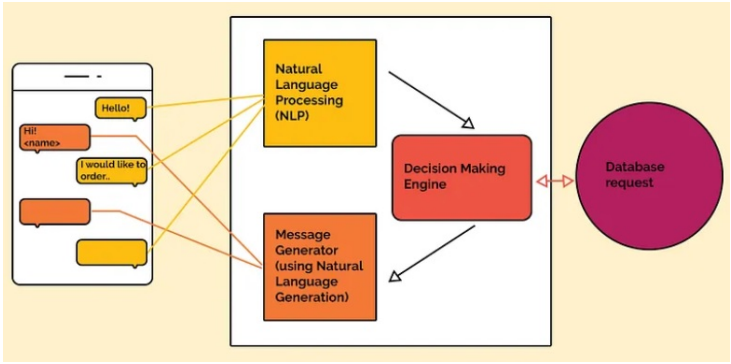

Figura 2.10 – Exemplo de NLG através de um chatbot.

Fonte: Medium (2023)

- **Classificação de entidades nomeadas** no PLN atribui automaticamente categorias a documentos, sendo vital para a organização e extração de informações de dados textuais extensos. Iniciando com o pré-processamento, envolvendo etapas como remoção de *stop words* e *tokenização*, o texto é vetorizado para representação numérica. A escolha do modelo, seja tradicional ou avançado como os modelos Transformer, é crítica. O treinamento, com dados rotulados, ajusta parâmetros para reconhecer padrões, sendo avaliado em dados não vistos. Esta técnica tem aplicações variadas, como categorização de notícias e filtragem de spam. Gienapp, Kircheis, Sievers et al. (2023) destaca a necessidade de aprimorar a capacidade de classificação em grandes conjuntos textuais, especialmente em ambientes governamentais. Essas perspectivas destacam a relevância contínua e oportunidades na classificação de texto com PLN. A Figura 2.11 exemplifica um processo de classificação com o modelo Transformer BERT. Esta é a técnica que será utilizada neste trabalho.

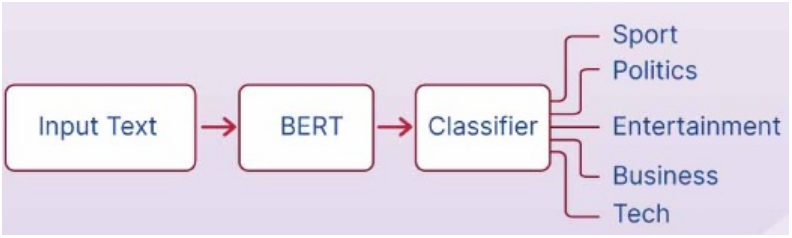

Figura 2.11 – Modelo genérico de classificação de texto utilizando modelo BERT.

Fonte: Harsha (2023)

2.3 Transformers

Os Transformers são modelos de *deep learning* que utilizam técnicas matemáticas de autoatenção para contextualizar conjuntos de dados de entrada em tarefas de PLN. Diferentemente de outras arquiteturas, como Redes Neurais Recorrentes (RNNs) e Redes Neurais Convolucionais (CNNs), os Transformers destacam-se pela capacidade de computa-

ção paralela, possibilitando treinamento mais rápido dos modelos em comparação com o processamento sequencial, aplicado nos outros modelos. A arquitetura Transformer descrita em Vaswani et al. (2017), Figura 2.12, foi introduzida em 2017 pelo grupo de inteligência artificial e *deep learning* do Google.

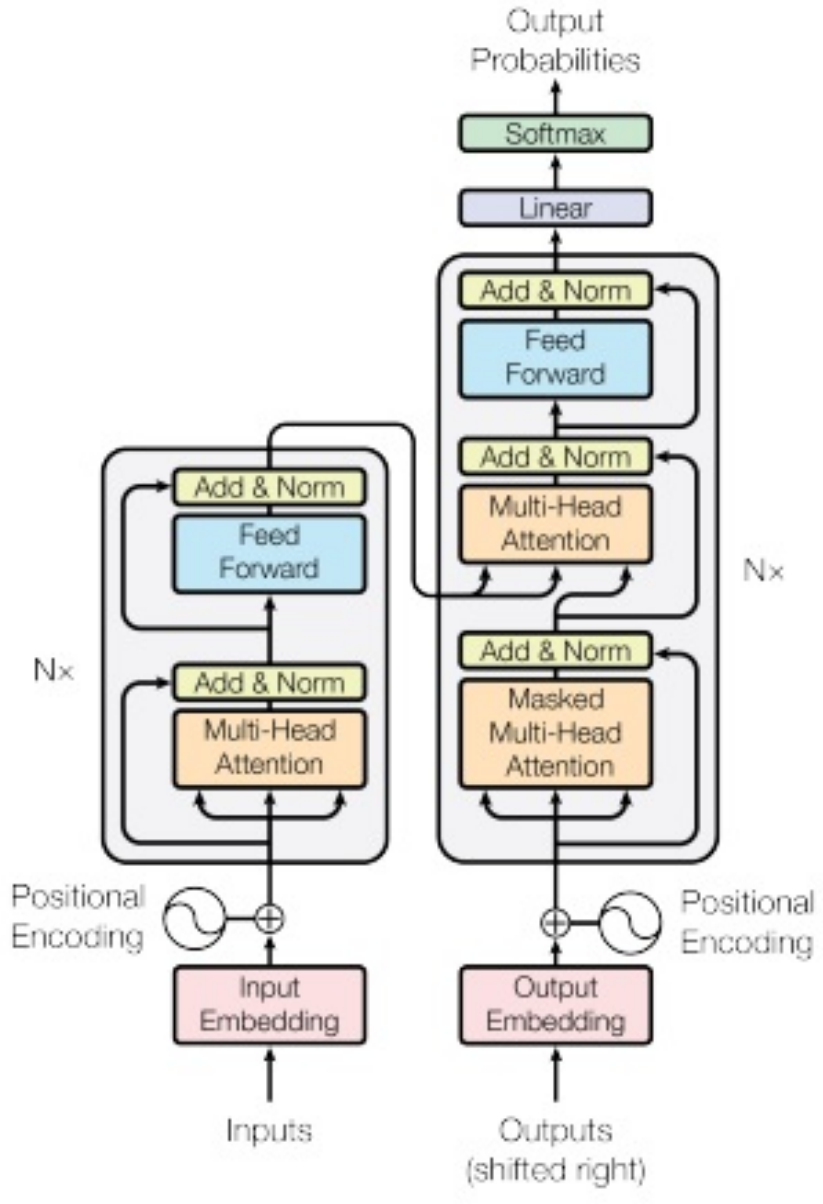

Figura 2.12 – O Transfomer - Arquitetura do Modelo.

Fonte: Vaswani et al. (2017)

De acordo com Daniel Jurafsky (2023), essa estrutura inovadora incorpora dois mecanismos principais: autoatenção e codificações posicionais. Esses mecanismos aprimoram a conexão entre palavras distantes, resultando em uma significativa melhoria na compreensão do contexto global do texto. Antes do advento dos Transformers, as RNNs eram predominantes em tarefas de PLN e lidavam com informações distantes, mas a inovação-chave do

Transformer é a capacidade de paralelização, permitindo o treinamento eficiente de modelos com grandes volumes de dados em menos tempo.

Em resumo, os Transformers mapeiam sequências de vetores de entrada (x_n) para sequências de vetores de saída (y_n). Esses modelos são compostos por pilhas de blocos de Transformer, cada bloco combinando camadas de normalização, redes feedforward, camadas de autoatenção e conexões residuais, conforme ilustrado na Figura 2.13.

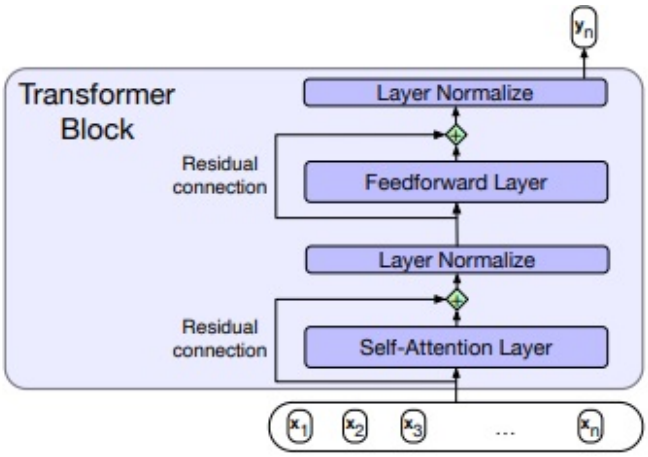

Figura 2.13 – Um bloco Transformer e suas camadas.

Fonte: Daniel Jurafsky (2023)

2.3.1 Redes Feedforward

Uma rede neural *feedforward* é um tipo de rede neural artificial simples composta por múltiplas camadas, sendo caracterizada pela ausência de ciclos em suas conexões, conforme descrito por Daniel Jurafsky (2023). Essas redes possuem três tipos de nós: nós de entrada, nós ocultos e nós de saída. A Figura 2.14 demonstra um exemplo simplificado de rede *feedforward*.

A camada de entrada, denotada por x, é um vetor de valores escalares simples, representando as entradas da rede. A parte central da rede é a camada oculta, composta por unidades ocultas h_i, cada uma realizando uma combinação linear ponderada de suas entradas, seguida pela aplicação de uma função de ativação não linear, como a função sigmoide, tangente hiperbólica ou ReLU.

Na arquitetura padrão, cada camada é totalmente conectada, o que significa que cada unidade em uma camada recebe como entrada as saídas de todas as unidades na camada anterior. Os parâmetros da camada oculta, como os pesos das conexões e os viéses, são representados de maneira eficiente por uma matriz de pesos W e um vetor de viés b. Cada elemento W_{ji} da matriz de pesos W representa o peso da conexão da i-ésima unidade de entrada x_i para a j-ésima unidade oculta h_j.

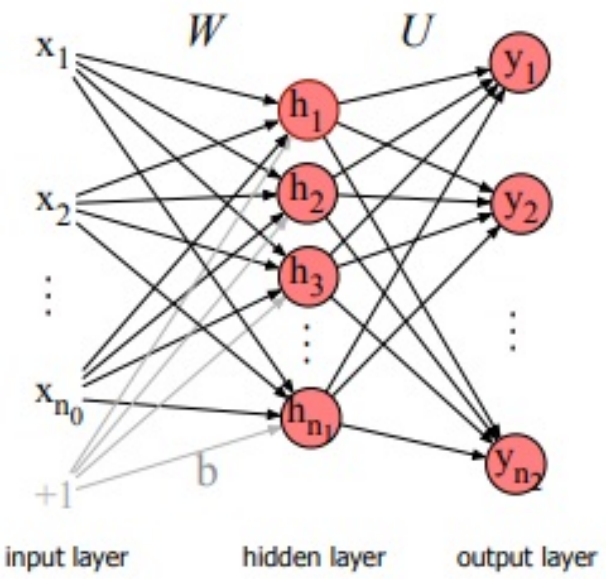

Figura 2.14 – Rede neural *feedforward* simples de duas camadas.

Fonte: Daniel Jurafsky (2023)

O calculo na camada oculta ocorre em três etapas: multiplicação da matriz de pesos pelo vetor de entrada x, adição do vetor de viés b, e aplicação da função de ativação. Utilizando a função sigmoide, σ, como exemplo, a saída da camada oculta h é dada de acordo com a Equação (2.1).

$$h = \sigma(Wx + b) \tag{2.1}$$

Assim, uma rede *feedforward* realiza uma propagação unidirecional da informação da camada de entrada através das camadas ocultas até a camada de saída, sem feedback das saídas para as camadas anteriores.

O papel da camada de saída é calcular um vetor y que proporciona uma distribuição de probabilidade entre os nós de saída, de acordo com camada oculta h. A transformação ocorre da seguinte maneira: assim como na camada oculta, a camada de saída possui uma matriz de pesos U, onde z, o vetor intermediário de saída, é obtido multiplicando U pelo vetor de entrada h, Equação (2.2).

$$z = Uh \tag{2.2}$$

O vetor z é uma representação de valores reais, mas para fins de classificação, precisamos de uma distribuição de probabilidades. A função *softmax* é utilizada para normalizar o vetor de valores reais, transformando-o em uma distribuição de probabilidade, onde todos os valores estão entre 0 e 1 e a soma é igual a 1. A função *softmax* é definida para um vetor z de dimensionalidade d de acordo com a Equação (2.3).

$$\text{softmax}(z_i) = \frac{exp(z_i)}{\sum_{j=1}^{d} exp(z_j)} \quad \text{para } 1 \leq i \leq d \tag{2.3}$$

Dessa forma resumida, a informação da camada oculta h é transformada através da multiplicação pela matriz de pesos U e normalizada pela função *softmax* para produzir um vetor y que representa uma distribuição de probabilidades entre os nós da camada de saída.

2.3.2 Autoatenção

As camadas de autoatenção representam a inovação chave nos Transformers. Daniel Jurafsky (2023) explica que essa técnica permite que uma rede analise e utilize informações de contextos amplos de maneira direta, sem depender de conexões recorrentes intermediárias. Em uma camada de autoatenção, cada elemento de entrada é processado com acesso a todos os elementos anteriores, resultando em uma saída independente. O processo envolve comparações entre elementos, representadas por pontuações normalizadas, que são utilizadas para calcular a distribuição de probabilidade. Essa distribuição é então empregada em uma soma ponderada, constituindo a saída da autoatenção.

A Figura 2.15 ilustra o fluxo de informação em uma única camada de autoatenção causal, ou seja, voltada para trás. Logo, no cerne da abordagem de autoatenção está a capacidade de comparar um item de interesse com uma coleção de outros itens, demonstrando sua relevância no contexto atual. O resultado dessas comparações é então usado para calcular uma saída para a entrada atual. Por exemplo, na Figura 2.15, o cálculo de y_3 é baseado em comparações entre a entrada x_3 e seus elementos precedentes x_1 e x_2. A abordagem utilizada para comparação entre elementos em uma camada de autoatenção é um produto escalar, e define a Equação de pontuação (2.4).

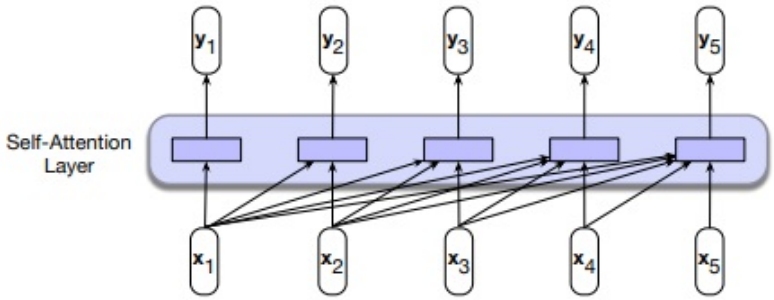

Figura 2.15 – Fluxo de informação em um modelo de autoatenção casual.

Fonte: Daniel Jurafsky (2023)

$$score(x_i, x_j) = x_i \cdot x_j \tag{2.4}$$

O resultado do produto escalar gera um valor entre $-\infty$ a ∞, sendo que quanto maior o valor, mais similares são os vetores comparados. Para o calculo de y_3, primeiramente

são computadas três pontuações: $x_3 \cdot x_1$, $x_3 \cdot x_2$ e $x_3 \cdot x_3$. Em seguida, as pontuações são normalizadas usando a função softmax para criar um vetor de pesos, α_{ij}, que indica a relevância proporcional de cada entrada para o elemento de entrada i, que é o foco de atenção atual, como demonstra a Equação (2.5).

$$\alpha_{ij} = \text{softmax}(\text{score}(x_i, x_j)) = \frac{\exp(\text{score}(x_i, x_j))}{\sum_{k=1}^{i} \exp(\text{score}(x_i, x_k))} \quad \forall j \leq i \tag{2.5}$$

Dadas as pontuações proporcionais em α, é gerado um valor de saída y_i somando os inputs precedentes, ponderados pelos seus respectivos valores de α, de acordo com a Equação (2.6).

$$y_i = \sum_{j \leq i} \alpha_{ij} x_j \tag{2.6}$$

A camada de autoatenção em um Transformer funciona de maneira análoga ao que foi previamente explicado. Nesse contexto, cada entrada desempenha uma das três funções no processo de atenção:

- Query (consulta): a entrada atual, ou seja, o foco do processo de atenção, comparada com as entradas anteriores.
- Key (chave): a entrada anterior comparada com a entrada atual.
- Value (valor): o valor utilizado para calcular a saída correspondente à entrada atual.

Isso implica na introdução das matrizes de pesos W_Q, W_K e W_V, que projetam cada vetor de entrada x_i em uma representação de sua função como consulta, chave ou valor, Equações (2.7).

$$q_i = W_Q x_i; \quad k_i = W_K x_i; \quad v_i = W_V x_i \tag{2.7}$$

A comparação entre a entrada atual x_i e uma entrada anterior x_j é realizada pelo produto escalar entre seu vetor de consulta q_i e o vetor de chave k_j do elemento anterior, conforme a equação (2.8).

$$\text{score}(x_i, x_j) = \frac{q_i \cdot k_j}{\sqrt{d_k}} \tag{2.8}$$

Aqui, d_k é a dimensão dos vetores de consulta e chave, e a divisão pelo fator $\sqrt{d_k}$ evita problemas numéricos e a perda efetiva de gradientes durante o treinamento. Como

na equação anterior, essa pontuação passa pela função softmax, resultando na saída atual calculada por uma soma ponderada, descrita pela Equação (2.9).

$$y_i = \sum_{j \leq i} \alpha_{ij} v_j \tag{2.9}$$

Neste caso, os pesos são determinados pelo vetor de valor v. Uma representação visual desse processo utilizando o modelo de autoatenção em um Transformer é mostrada na Figura 2.16, semelhante ao exemplo apresentado na Figura 2.15. De acordo com Daniel

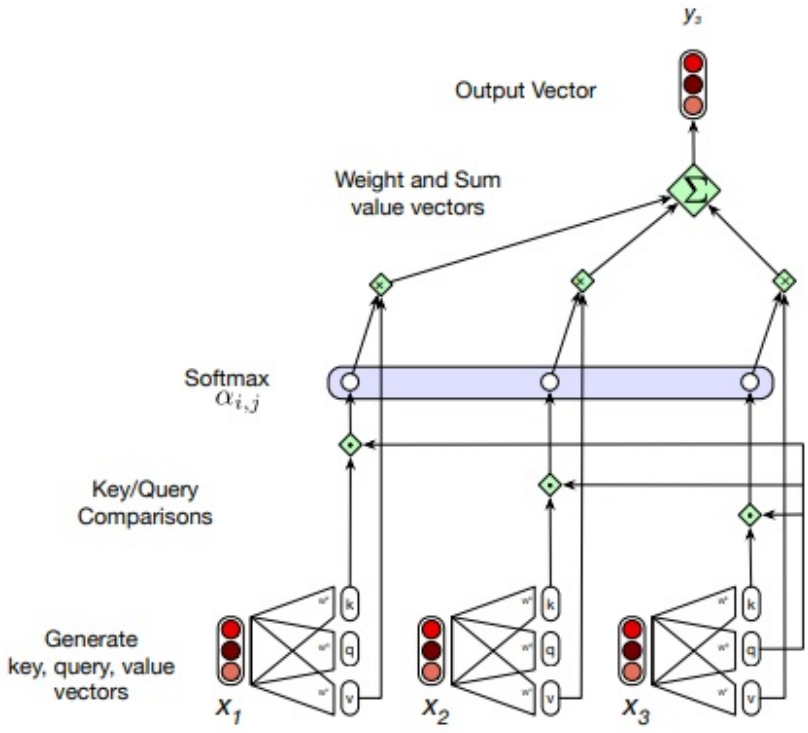

Figura 2.16 – Fluxo de informação em um modelo de autoatenção em um Transformer.

Fonte: Daniel Jurafsky (2023)

Jurafsky (2023), o processo de autoatenção explicado acima, é destinado ao cálculo de um único valor de saída, podendo ser paralelizado devido à independência do cálculo de cada saída. Utilizando matrizes para representar todas as entradas em uma matriz X, de acordo com as Equações em (2.10).

$$Q = XW_Q; \quad K = XW_K; \quad V = XW_V \tag{2.10}$$

Ao expressar as Equações (2.8) e (2.9) em termos de matrizes e combiná-las, o processo de autoatenção pode ser simplificado para a equação (2.11).

$$\text{AutoAtenção}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.11}$$

Devido ao termo QK^T calcular não apenas as comparações das entradas anteriores, mas também das entradas subsequentes, os elementos da porção triangular superior da matriz QK^T são computados como $-\infty$, eliminando qualquer conhecimento das palavras seguintes na sequência.

2.3.3 Conexões Residuais

Em redes profundas, conexões residuais referem-se à transferência de informações de uma camada inferior para uma camada superior sem passar pela camada intermediária. Essa abordagem, introduzida por He et al. (2016), melhora o aprendizado ao permitir que as informações da ativação avancem e que o gradiente retroceda, pulando uma camada e proporcionando às camadas de nível superior acesso direto às informações das camadas inferiores.

Nos Transformers, as conexões residuais são implementadas adicionando o vetor de entrada de uma camada ao vetor de saída antes de prosseguir. No bloco de Transformer apresentado na Figura 2.13, as conexões residuais são utilizadas tanto nas subcamadas de autoatenção quanto nas subcamadas feedforward. Esses vetores somados são então normalizados por camada, conforme proposto por Ba, Kiros e Hinton (2016). Assim, considerando uma camada como um vetor, a função resultante calculada em um bloco de Transformer pode ser expressa pelas Equações (2.12) e (2.13).

$$z = \text{LayerNorm}(x + \text{SelfAttention}(x)) \tag{2.12}$$

$$y = \text{LayerNorm}(z + \text{FFN}(z)) \tag{2.13}$$

2.3.4 Normalização

De acordo com Daniel Jurafsky (2023), a normalização por camada é uma técnica utilizada para aprimorar o desempenho do treinamento em redes neurais profundas, mantendo os valores de uma camada oculta em uma faixa que facilita o treinamento baseado em gradientes. O primeiro passo na normalização por camada é determinar a média, μ, e o desvio padrão, σ, sobre os elementos do vetor a ser normalizado. Dada uma camada oculta com dimensionalidade d_h, esses valores são calculados pela Equações (2.14) e (2.15).

$$\mu = \frac{1}{dh}\sum_{i=1}^{dh} x_i \tag{2.14}$$

$$\sigma = \sqrt{\frac{1}{dh}\sum_{i=1}^{dh}(x_i - \mu)^2} \tag{2.15}$$

Com esses valores, os componentes do vetor são normalizados subtraindo a média de cada um e dividindo pelo desvio padrão. O resultado dessa computação é um novo vetor com média zero e desvio padrão de um, de acordo com a Equação (2.16).

$$\hat{x} = \frac{(x - \mu)}{\sigma} \tag{2.16}$$

Finalmente, na implementação padrão da normalização por camada, são introduzidos dois parâmetros aprendíveis, γ e β, representando os valores de ganho e deslocamento, respectivamente. Esses parâmetros desempenham um papel crucial no processo de normalização e se relacionam com o vetor normalizado pela Equação (2.17).

$$\text{LayerNorm} = \gamma\hat{x} + \beta \tag{2.17}$$

2.3.5 Codificação Posicional

Em um modelos de Transformer, representar a posição de cada token na sequência de entrada é um aspecto crucial, ao contrário das RNNs tradicionais, onde essa informação é inerentemente codificada na estrutura do modelo. Os Transformers não possuem uma compreensão embutida das posições relativas ou absolutas dos tokens na entrada, e foi necessário introduzir o conceito de codificação posicional. Conforme descrito por Daniel Jurafsky (2023), a solução envolve a incorporação das posições dos tokens nos próprios tokens, por meio da soma de codificações posicionais, como ilustra a Figura 2.17.

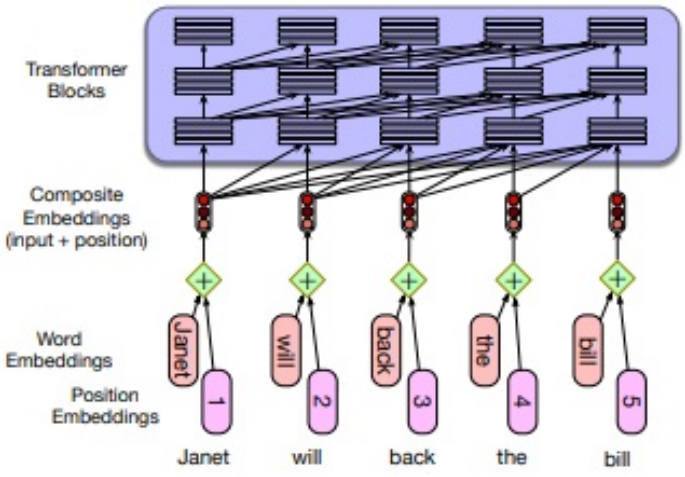

Figura 2.17 – Modelo de codificação posicional absoluta.

Fonte: Daniel Jurafsky (2023)

Assim, a codificação de entrada resultante, que inclui informações tanto de tokens quanto de posição, são apreendidas juntamente com outros parâmetros do modelo, durante o treinamento do mesmo.

Diversas opções de codificações posicionais, sejam aprendidas ou fixas, estão disponíveis. O modelo Transformer emprega padrões estáticos, como sinais senoidais e cossenoidais, para codificar a posição dos tokens, sendo eficaz em cenários de dados limitados. Alternativamente, as representações posicionais relativas consideram a relevância da posição relativa ao calcular a codificação posicional de um token. A abordagem de codificação posicional rotativa combina elementos de codificação posicional absoluta e relativa, destacando-se em várias tarefas, como exemplificado no GPT-Neo, segundo Lewis Tunstall Leandro von Werra (2021).

No trabalho de Vaswani et al. (2017), que introduziu o modelo de Transformer, funções seno e cosseno de diferentes frequências são empregadas para facilitar o aprendizado das posições relativas pelo modelo. De acordo com Daniel Jurafsky (2023), apesar do bom desempenho dessas funções, o aprimoramento contínuo das representações de posição permanece um tema de pesquisa ativo.

2.4 Métricas de Avaliação

Para avaliar o desempenho de um modelo e a qualidade dos resultados que ele produz, geralmente são utilizadas quatro métricas principais: acurácia, precisão, sensibilidade e o F-Score. Contudo, antes de abordar essas métricas, de forma mais específica, é necessário compreender sobre a matriz de confusão gerada pelo modelo. Considerando um modelo de classificação binário, ela é uma ferramenta valiosa que organiza os resultados das predições em quatro categorias:

- **Verdadeiro-positivo (VP):** São casos em que o modelo identifica corretamente as entidades que fazem parte do conjunto que precisa ser reconhecido.
- **Verdadeiro-negativo (VN):** Refere-se aos casos em que o modelo não identifica entidades que realmente não deveriam ser reconhecidas.
- **Falso-positivo (FP):** São casos em que o modelo erroneamente identifica entidades que não deveriam ser identificadas.
- **Falso-negativo (FN):** Indica situações em que o modelo falha em identificar entidades que deveriam ter sido reconhecidas.

Para completar, Andreas C. Müller (2016) demonstra uma matriz de confusão genérica, Figura 2.18, e define as amostras corretamente classificadas pertencentes à classe positiva e classe negativa. No entanto, não é usual avaliar um modelo através de sua matriz de confusão, embora seja possível obter muitas percepções ao observar todos os seus aspectos, o processo tende a ser manual e qualitativo. As métricas de acurácia, precisão, sensibilidade e

o F-Score surgem como uma abordagem matemática para resumir as informações na matriz de confusão.

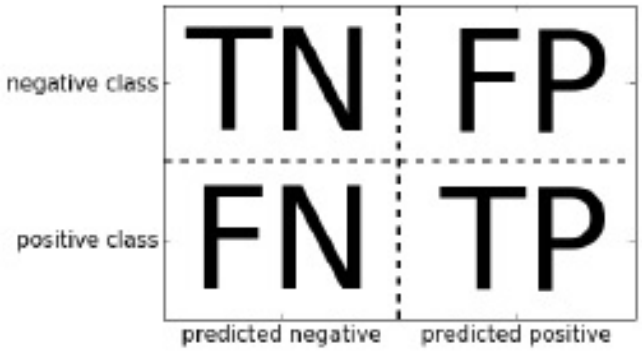

Figura 2.18 – Matriz de confusão para classificação binária.

Fonte: Andreas C. Müller (2016)

A acurácia fornece uma visão geral da precisão do modelo ao mensurar a proporção de predições corretas em relação ao total de predições. É a métrica mais simples e amplamente utilizada, ela é calculada de acordo com a Equação (2.18).

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.18}$$

No entanto, conforme descrito por Andreas C. Müller (2016), sua aplicação isolada pode ser enganosa em cenários de desbalanceamento de classe. Em situações em que há uma disparidade significativa entre o número de instâncias de cada classe, a acurácia pode ser influenciada, levando a uma interpretação inadequada do desempenho do modelo. O desbalanceamento de classe pode distorcer a acurácia, pois um modelo pode alcançar uma alta taxa de acertos simplesmente prevendo predominantemente a classe majoritária.

Ainda de acordo com Andreas C. Müller (2016), a precisão mede quantas das amostras previstas como positivas são realmente positivas, é calculada pela Equação (2.19):

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{2.19}$$

É usada como métrica de desempenho quando o objetivo é limitar o número de falsos positivos. Portanto, é uma métrica de maior valor quando se deseja um modelo que não deve produzir muitos falsos positivos, devido a algum risco associado, ou seja, que tenha alta precisão. A precisão também é conhecida como valor preditivo positivo (VPP).

Já a sensibilidade, por outro lado, mede quantas amostras positivas são capturadas pelas previsões positivas, Equação (2.20):

$$\text{Sensibilidade} = \frac{VP}{VP + FN} \tag{2.20}$$

A sensibilidade é usada como métrica de desempenho quando é necessário identificar todas as amostras positivas, ou seja, quando é importante evitar falsos negativos. Ela também é chamada de taxa de acerto ou taxa de verdadeiro positivo (TVP).

Embora a precisão e a sensibilidade sejam medidas muito importantes, analisá-las separadamente não fornece uma visão completa sobre a avaliação de um modelo de classificação. Logo o F-Score, surgiu como uma medida que incorpora ambas as métricas, como demonstra a Equação (2.21):

$$F_\beta = \frac{(\beta^2 + 1) \cdot \text{Precisão} \cdot \text{Sensiblidade}}{(\beta^2) \cdot (\text{Precisão} + \text{Sensiblidade})} \tag{2.21}$$

Conforme explicado por Daniel Jurafsky (2023), o parâmetro β é responsável por ponderar a relevância da precisão e da sensibilidade. Se β for maior que 1, a sensibilidade receberá uma ênfase maior, enquanto para valores de β menores que 1, a precisão terá uma importância superior. Contudo, o valor mais comumente adotado é $\beta = 1$, indicando uma igual ponderação das duas medidas, este é o valor utilizado neste trabalho. Portanto, a expressão da Equação (2.21) é ajustada para a Equação (2.22), segundo Andreas C. Müller (2016), uma média harmônica da precisão e sensibilidade:

$$F_1 = \frac{2 \cdot \text{Precisão} \cdot \text{Sensiblidade}}{(\text{Precisão} + \text{Sensiblidade})} \tag{2.22}$$

F1-Score é uma métrica complementar a acurácia, e de maior confiabilidade, para avaliação de modelos treinados em conjuntos dados com desequilíbrio de classe.

2.4.1 Métricas para Classificação Multiclasse

Após discutir a avaliação de tarefas de classificação binária, é essencial compará-la com as métricas usadas na avaliação da classificação multiclasse, que é empregada neste trabalho. Basicamente, todas as métricas para classificação multiclasse são derivadas de métricas de classificação binária, mas calculadas em média sobre todas as classes, de acordo com Andreas C. Müller (2016). A acurácia para classificação multiclasse é novamente definida como a fração de exemplos classificados corretamente. E, novamente, quando as classes estão desequilibradas, a acurácia não representa uma boa medida de avaliação, de forma única.

Para avaliação multiclasse, é gerada uma matriz de confusão mais extensa, permitindo a criação de uma tabela com valores de precisão, sensibilidade e F1-Score. A abordagem das métricas multiclasse consiste em calcular a avaliação binária para cada classe, considerando essa classe como positiva e as demais como negativas. A média dessas pontuações por classe é então calculada usando uma das seguintes estratégias:

- A média **"micro"** avalia o número total de falsos positivos, falsos negativos e verdadeiros positivos em todas as classes, de forma global, calculando precisão, sensibilidade e F1-Score com base nessas contagens. Recomendada quando cada amostra é igualmente relevante, independentemente da classe.
- A média **"macro"** calcula pontuações não ponderadas por classe, atribuindo igual peso a todas as classes. Indicada quando cada classe é igualmente relevante, independentemente do número de instâncias.
- A média **"weighted"** pondera as métricas com base no desequilíbrio de amostras entre as classes. Classes com mais amostras têm mais peso na média final, garantindo uma avaliação que reflete a influência proporcional de cada classe no desempenho global do modelo. Essa técnica é valiosa em casos de desigualdade na distribuição de amostras entre as classes, sendo empregada neste trabalho.

3 Trabalhos Relacionados

Este capítulo visa realizar uma revisão detalhada da literatura no âmbito do processamento de linguagem natural para classificação de textos. Para alcançar esse objetivo, foram examinados trabalhos relacionados, a fim de posicionar este estudo dentro do contexto científico, ao mesmo tempo em que se busca identificar as técnicas mais recentes e relevantes aplicadas nessa área. Essa revisão visa proporcionar uma compreensão abrangente do estado atual da tarefa de classificação de texto e das abordagens mais promissoras que têm sido exploradas, que serviram de base para o desenvolvimento deste trabalho.

3.1 Classificação de entidades textuais

Ao longo dos últimos anos, o campo de processamento de linguagem natural tem testemunhado avanços significativos, impulsionados principalmente pela ascensão de modelos baseados em Transformers, notavelmente o BERT e sua contraparte mais leve, DistilBERT. Esses modelos pré-treinados têm se destacado em diversas tarefas de PLN, incluindo a classificação de textos.

O trabalho de Vaswani et al. (2017), precursor dos modelos Transformers, delineou a arquitetura original e sua aplicação em tarefas de tradução automática. A capacidade dos Transformers de capturar relações semânticas complexas em diferentes línguas motivou pesquisadores a adaptar essa arquitetura para outras aplicações em vários domínios. Posteriormente, Devlin et al. (2018) introduziu o BERT, um modelo capaz de capturar relações semânticas complexas em contextos bidirecionais. Sua aplicação em tarefas de classificação textual tem gerado resultados notáveis, proporcionando uma compreensão mais profunda do contexto e das relações semânticas presentes nos textos. Em paralelo, Sanh et al. (2019) propuseram o DistilBERT, uma versão mais eficiente do BERT que mantém um desempenho sólido, tornando-o mais acessível computacionalmente. A adaptação desses modelos para a classificação de entidades textuais tem sido objeto de investigação em vários estudos, Howard e Ruder (2018), por exemplo, exploraram o uso de *embeddings* contextuais para melhorar a representação semântica em tarefas de classificação, mostrando a capacidade desses modelos em lidar com a complexidade linguística de diferentes domínios.

Daniel Jurafsky (2023) estabelece uma base teórica robusta abordando métricas convencionais amplamente empregadas em estudos científicos, tais como precisão, sensibilidade e F1-score, e destaca suas aplicações específicas em tarefas de PLN. Além disso, ele explora estratégias de adaptação e transferência de conhecimento, visando otimizar o desempenho desses modelos em contextos particulares. Em complemento a essa abordagem, o trabalho de Howard e Ruder (2018) também enfatiza a importância de técnicas específicas de *fine-tuning*

para ajustar modelos pré-treinados em tarefas de classificação. Esse estudo oferece percepções valiosas para a implementação prática dessas abordagens, contribuindo de maneira significativa para o aprimoramento do desempenho em contextos mais especializados. Essas considerações conjuntas enriquecem a compreensão e a aplicação das métricas e estratégias discutidas, promovendo uma visão mais abrangente e coesa no contexto de treinamento de modelos envolvendo linguagem natural.

No contexto de dados públicos, os modelos de Transformer possuem distintas aplicações que apresentam soluções para tarefas governamentais. Limsopatham (2021), por exemplo, apresentou um artigo que investiga a eficácia do BERT na classificação de documentos legais, com objetivo de treinar este modelo para o contexto jurídico, considerando suas características únicas. Ele utilizou os conjuntos de dados *ECHR Violation* e *Overruling Task*, para representar tarefas multirrótulo e binária, respectivamente, o estudo compara técnicas de adaptação para lidar com documentos extensos. Os resultados obtidos são demonstrados na Tabela 3.1. Foi obtido boa performance para a tarefa de classificação binária, enquanto a tarefa de classificação multirótulo destaca a necessidade de adaptação explícita do BERT para alcançar desempenho eficaz em documentos jurídicos longos, sugerindo que o pré-treinamento em documentos similares à tarefa alvo, abordado por outros modelos, melhora o desempenho em diferentes cenários.

Base de Dados	**Modelo**	**F1-Ccore**
ECHR Violation	BigBird	73,08%
	ECHR-Legal-BERT	72,13%
	LongFormer	72,38%
	BERT	71,10%
	Harvard-Law-BERT	70,10%
	RoBERTa	70,00%
Overruling Task	BERT	96,56%
	ECHR-Legal-BERT	97,25%
	Harvard-Law-BERT	97,56%
	RoBERTa	96,83%
	BigBird	95,70%
	LongFormer	95,69%

Tabela 3.1 – Comparação de desempenho nas bases de dados ECHR Violation e Overruling Task, para tarefas de classificação multirótulo e binária, respectivamente.

No mesmo contexto, Vatsal, Meyers e Ortega (2023) também utilizou modelos de Transformer para realizar tarefas semelhantes de classificação de documentos extensos, especialmente em decisões da Suprema Corte dos Estados Unidos. Aqui, também nota-se uma complexidade ao aplicar modelos BERT de maneira direta em documentos longos, diferentes técnicas de classificação foram abordadas para a tarefa de classificação multirrótulo: para 15 categorias amplas e outra com 279 categorias mais detalhadas. Foram obtidos bons resultados, apresentando uma precisão de 80% nas 15 categorias amplas e 60% nas

279 categorias detalhadas. Os melhores resultados, associados a métricas, deste trabalho são representados na Tabela 3.2. Essas descobertas contribuem para a compreensão dos

Nº de Categorias	Modelo	Acurácia	Precisão	F1-Score
15	BERT	0.774	0.777	0.771
	RoBERTa	0.779	0.785	0.778
	Legal-BERT	0.801	0.805	0.800
279	BERT	0.563	0.514	0.519
	RoBERTa	0.536	0.465	0.479
	Legal-BERT	0.609	0.584	0.575

Tabela 3.2 – Comparação de métricas de desempenho para modelos com 15 e 279 categorias, treinado sobre publicações da suprema corte americana.

desafios específicos ao usar BERT em documentos longos, oferecendo avanços significativos no desempenho de classificação.

Além disso, a pesquisa de Brown et al. (2020) destacou a utilidade dos Transformers na classificação de textos em dados públicos para a detecção de notícias falsas. A capacidade desses modelos em compreender nuances semânticas e contextuais foi explorada para identificar padrões que indicam a veracidade ou falsidade das informações. Outro cenário explorado é a classificação de documentos governamentais abordada por Gowrisankar e Thing (2023) onde foram empregadas técnicas para categorizar documentos públicos, fornecendo uma estrutura robusta para a organização e recuperação de informações críticas. Já Büyüköz, Hürriyetoğlu e Özgür (2020) adaptaram o DistilBERT para categorização automática de notícias governamentais. Esses estudos não apenas demonstraram a versatilidade desses modelos, mas também indicaram sua capacidade de se adaptar a diferentes contextos e tipos de documentos. No desenvolvimento de conjuntos de dados rotulados, Liang et al. (2018) contribuíram com estratégias eficazes para garantir representatividade e diversidade nas amostras, elementos cruciais para o treinamento e avaliação robusta de modelos de PLN.

Ao analisar estudos similares, observa-se uma significativa melhoria na precisão e eficiência na classificação de entidades textuais. Modelos baseados em Transformers, como BERT e DistilBERT, demonstram habilidade para lidar com a diversidade linguística e temática em diversos dados, tornando-os escolhas promissoras para categorização em larga escala. A revisão destes trabalhos destacam a eficácia desses modelos na classificação multiclasse de textos públicos, ressaltando também a importância de considerar características específicas de documentos governamentais brasileiros. Os resultados desses estudos fornecem uma base sólida para a implementação prática e desenvolvimento do modelo proposto, oferecendo percepções valiosas a pesquisadores e profissionais interessados na aplicação dessas tecnologias em contextos semelhantes.

4 Metodologia

Com o intuito de criar uma base de publicações consistente, relativas a licitações, foram elaboradas técnicas que combinam processos de aquisição de dados, tratamento e armazenamento para o posterior treinamento de aprendizado de máquina. O projeto foi subdividido em cinco fases distintas, cada uma desempenhando um papel necessário para que os objetivos desse trabalho fossem alcançados.

1. **Mapeamento das entidades:** Nesta etapa, as principais entidades e informações relevantes foram identificadas nas publicações de licitações presentes no portal de compras governamentais, disponível em API de Compras Governamentais (2023). Essa etapa envolveu a definição clara de campos que seriam extraídos para posterior análise.

2. **Aquisição das publicações de licitações:** As publicações do site de compras governamentais foram coletadas e armazenadas, abrangendo uma ampla gama de dados, contento todas informações relevantes aos processos licitatórios. Essa aquisição permitiu a criação de um conjunto de dados representativo numericamente e diversificado em relação as entidades governamentais.

3. **Enriquecimento da base de dados com publicações do DOU e criação de rótulos:** Foi realizado um procedimento de agregação da base anterior com publicações correlacionadas do DOU, decorrente de um banco de publicações da ferramenta *deep vacuity*, utilizada no trabalho de Marcos Cavalcanti Lima (2021), para cada publicação obtida na fase anterior. Este processo possibilitou a associação publicações de duas naturezas distintas, associadas a processos licitatórios, permitindo a criação de uma base de dados mais complexa para posterior utilização do treinamento dos modelos de Transformer.

4. **Validação:** Para garantir a integridade dos dados e a correspondência precisa entre as publicações agregadas na etapa anterior, um processo de validação foi necessário. Esta etapa incluiu a verificação de consistência, detecção de possíveis erros e refinamento dos dados.

5. **Treinamento de modelos de Transformers para a tarefa de classificação:** A fase final envolveu o treinamento de modelos de Transformers, com o objetivo de classificar as publicações presentes na base de dados. Os modelos foram preparados para reconhecer padrões, associar informações e agrupar entidades textuais de acordo com os critérios predefinidos.

A Figura 4.19 apresenta resumidamente a ação, o objetivo e os resultados esperados em cada etapa proposta, que serão detalhadas nas próximas seções.

	MAPEAMENTO	AQUISIÇÃO	ENRIQUECIMENTO	VALIDAÇÃO	TREINAMENTO
AÇÃO	Realizar um mapeamento das entidades nas publicações de licitações.	Buscar e extrair as publicações de licitações da API de compras do governo federal.	Buscar e extrair publicações do DOU, de um banco de dados contendo uma coleção de publicações, com base nas entidades definidas nas etapas anteriores.	Validar as publicações da etapa anterior com base nas entidades mapeadas na primeira etapa.	Realizar o treinamento de dois modelos de Transformers: BERT e DistilBERT.
OBJETIVO	Identificar as entidades relevantes para a busca e validação de publicações.	Formar a primeira base de dados com publicações de licitações.	Formar uma base de dados enriquecida que relaciona publicações de processos licitatórios de naturezas distintas.	Garantir a consistência e o relacionamento das publicações de licitações para a base de dados	Produzir modelos capazes de classificar publicações de licitações em suas respectivas classes, com métodos que levam em consideração o contexto.
RESULTADOS	Uma lista de entidades para buscar publicações relacionadas as licitações e outra lista de entidades para valida-lás.	Base de dados de licitações adquirida do site de compras governamentais.	Base de dados de publicações enriquecida com publicações do DOU e rotulada.	Obter uma base de dados validada e consistente, com publicações correlacionadas rotuladas corretamente.	Modelos treinados, análise dos resultados obtidos e comparativo entre os modelos.

Figura 4.19 – Metodologia proposta.

Em conjunto, essas etapas proporcionam a criação de um *data pipeline* completo, compreendendo todo o ciclo, desde a coleta das publicações até sua classificação em rótulos específicos, conforme representado na Figura 4.20.

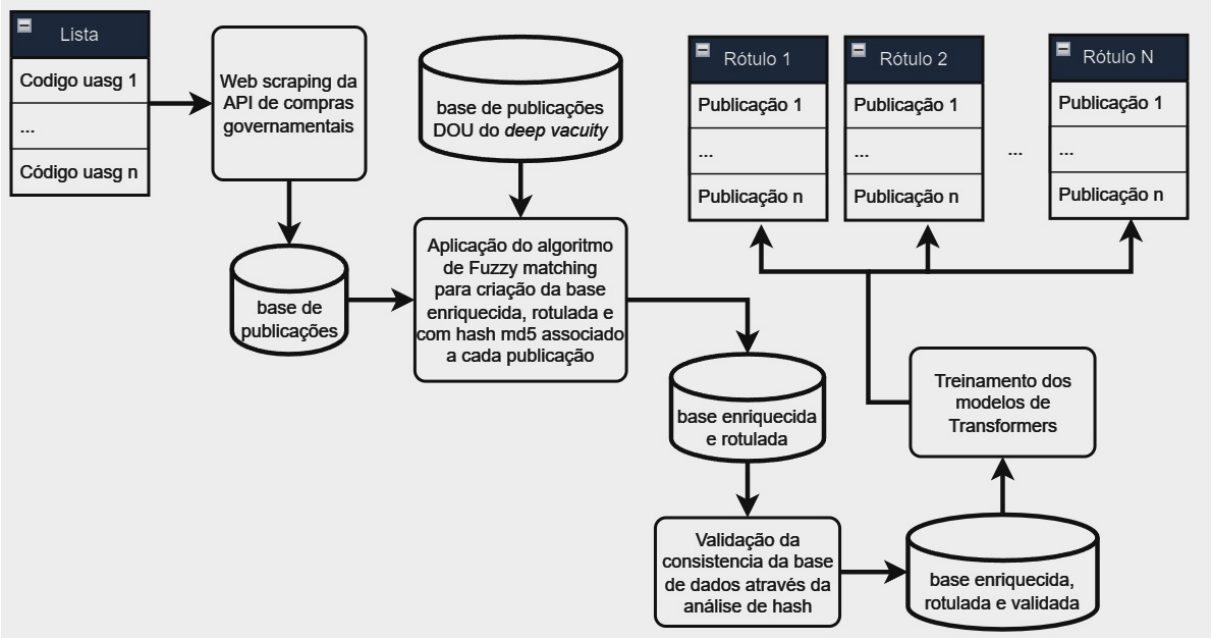

Figura 4.20 – Data pipeline do fluxo de trabalho desenvolvido.

4.1 Mapeamento das Entidades

A etapa inicial compreendeu a identificação das potenciais entidades associadas às licitações, com o propósito de avaliar sua relevância com base em critérios específicos. Esta etapa adquire relevância fundamental, visto que um mapeamento eficaz facilita as fases subsequentes de busca e validação das publicações. Para a realização dessa tarefa, o ponto de partida foi o dicionário de dados da API de Compras Governamentais (2023), que contém descrições de 20 entidades vinculadas a cada licitação, bem como exemplos de utilização da API que abrangem a obtenção de publicações licitatórias.

No contexto já abordado, foram identificadas 20 entidades ligadas a licitações, as quais estão enumeradas na Tabela 4.3. Durante o processo de identificação dessas entidades, uma amostra contendo uma série de informações relacionadas às entidades licitatórias foi utilizada. Dois fatores cruciais foram criteriosamente considerados: a unicidade e a frequência das entidades nas publicações. A abordagem destes dois fatores foi desenvolvida no escopo deste estudo para otimizar a eficácia da busca por publicações e solucionar desafios enfrentados durante a etapa de busca, os quais serão abordados em detalhes mais adiante.

Entidade	**Descrição**
	Fonte: API de Compras Governamentais (2023)
codigo do item no catalogo	Código do Item no Catálogo.
data abertura proposta	Data de Abertura da Proposta.
data entrega edital	Data de Entrega do Edital.
data entrega proposta	Data de Entrega da Proposta.
data publicacao	Data da Publicação da Licitação.
endereco entrega edital	Endereço de Entrega do Edital.
funcao responsavel	Função do Responsável pela Licitação.
identificador	Identificador da Licitação.
informacoes gerais	Informações Gerais.
modalidade	Código da Modalidade da Licitação.
nome responsavel	Nome do Responsável pela Licitação.
numero aviso	Número do Aviso da Licitação.
numero item licitacao	Número Item Licitação.
numero itens	Número de Itens.
numero processo	Número do Processo.
objeto	Objeto da Licitação.
situacao aviso	Situação do Aviso.
tipo pregao	Tipo do Pregão.
tipo recurso	Tipo do Recurso.
uasg	Código da UASG.

Tabela 4.3 – Descrição das entidades de uma licitação.

4.1.1 Unicidade

A unicidade das entidades desempenha um papel fundamental na busca por publicações relacionadas a licitações específicas. A presença de unicidade não apenas limita a quantidade de publicações recuperadas do DOU, mas também melhora a eficiência da busca, uma vez que a maior parte das publicações retornadas está diretamente ligada à licitação em consideração.

Realizar buscas usando entidades que carecem da característica de unicidade levaria a um grande volume de publicações que compartilham essas entidades, mas que não estão realmente associadas à licitação desejada. Um exemplo evidente de entidades com unicidade são o Identificador da Licitação e o Número do Processo, já que cada licitação é distinta e possui um identificador e número de processo exclusivos.

Por outro lado, a Unidade Administrativa de Serviços Gerais (UASG) não apresenta unicidade, uma vez que é compartilhada por diversas licitações e documentos públicos. Para mapear as entidades com base na unicidade, um procedimento foi aplicado à amostra de licitações:

1. Para cada entidade presente em cada licitação, buscas foram executadas nas publicações em Imprensa Nacional (2023), utilizando a própria entidade como critério de pesquisa.

2. Em seguida, para cada publicação identificada, foi verificado se ela de fato estava relacionada à licitação em questão. No caso de entidades que retornavam um alto número de publicações, cerca de milhares, considerou-se que a unicidade era baixa, uma vez que uma única licitação gera, em média, dezenas de publicações.

3. Com base no total de publicações recuperadas e na proporção de publicações verdadeiramente ligadas à licitação, foi possível chegar a um critério para avaliar o grau de unicidade de cada entidade de acordo com a porcentagem de publicações relacionadas sob a quantidade total de publicações retornadas.

4. Caso o critério de unicidade fosse satisfatório, as mesmas etapas eram validadas no banco de dados de publicações.

4.1.2 Frequência

A consideração da frequência das entidades nas publicações é outro fator necessário. Nesse contexto, a frequência não se refere à ocorrência da entidade em uma única publicação, mas sim à sua presença em múltiplas publicações relacionadas ao mesmo processo licitatório. Quanto mais frequente a entidade aparecer, mais eficiente será sua utilização na busca das publicações correspondentes à licitação. Para mapear as entidades com base na frequência, foi seguido o seguinte procedimento em uma amostra de licitações:

1. Para cada entidade de cada licitação, realizou-se uma busca por publicações em Imprensa Nacional (2023), usando a própria entidade como termo de pesquisa.

2. Em seguida, para cada publicação identificada, foi verificado se estava relacionada à licitação em questão. Para esse processo, foram utilizadas somente entidades que resultaram em um número razoável de publicações, na ordem das dezenas.

3. Após obter todas as publicações vinculadas à licitação, a frequência de cada entidade foi determinada com base na proporção de publicações que mencionam a entidade em seu conteúdo.

4. Se os critérios de frequência foram atendidos, as mesmas etapas foram aplicadas à base de dados de publicações.

É relevante destacar que a seleção das entidades com base nas características de unicidade e frequência é de importância crítica. Entidades que possuem unicidade, mas raramente são mencionadas nas publicações, têm menos relevância na busca por essas publicações. Da mesma forma, entidades frequentes, mas sem unicidade, tornam a busca ineficiente e, em alguns casos, impraticável. Logo, a seleção das entidades para busca e validação das publicações requer um equilíbrio sensato entre essas duas características.

4.1.3 Variações das Entidades

Além de mapear as entidades em termos de unicidade e frequência, a etapa de mapeamento também foi importante para compreender as diferentes formas pelas quais uma mesma entidade é mencionada nos textos das publicações. Para a entidade Número do Processo por exemplo, que é dada por XXXXX.YYYYYY/ZZZZ-KK, em que X, Y, Z e K são dígitos de 0 a 9. Pode-se observar que, nos textos das publicações pode aparecer no formato XXXXXYYYYYY/ZZZZKK, XXXXXYYYYYY/ZZZZ-KK ou XXXXX.YYYYYYZZZZKK. O mesmo ocorreu para a entidade Identificador da Licitação que foi observada de duas maneiras distintas: XXXXXX-Y-ZZZZZ-KKKK ou simplesmente XXXXXXYZZZZZKKKK, também em que X, Y, Z e K são dígitos de 0 a 9. Assim, para cada entidade, foram definidas variações considerando a inspeção manual de dezenas de publicações, o que se mostrou suficiente para esse propósito.

4.2 Aquisição das Publicações de Licitações

Com base no trabalho realizado na etapa anterior, sobre as entidades selecionadas e seus respectivos valores associados a cada licitação, e com a validação dos registros em publicações do DOU em Imprensa Nacional (2023), foi realizado a busca e armazenamento de dados de licitações em API de Compras Governamentais (2023) de forma automatizada,

utilizando como parâmetro de busca os códigos das Unidades Administrativas de Serviços Gerais. Nesta etapa, como a intenção foi preencher a primeira base de dados com o maior número de elementos possível, foi utilizada uma entidade de baixa unicidade e alta frequência para aquisição de maior quantidade de registros, de forma genérica. Conforme ilustrado no fluxograma da Figura 4.21, esta etapa consistiu em um laço de repetição responsável por enviar requisições a *API*. A estrutura de repetição é responsável por percorrer a lista de códigos de UASG, extraída do Catálogo de Unidades Administrativas de Serviços Gerais fornecido por COMPRANET (2023), realizar a requisição à *API* e armazenar os registros de licitações em um arquivo *CSV* e alguns metadados associados a requisição como o status da requisição e tempo de resposta. Por se tratar de uma *API* não há necessidade de técnicas robustas para ter acesso democratizado aos dados, porém para otimizar e maximizar a coleta de dados foi utilizada 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 4.21 – Fluxograma de aquisição dos dados de licitações.

4.2.1 Montagem da URL

Para realizar a pesquisa de dados de licitações na API de Compras Governamentais foi necessário um processo de montagem da URL para posterior requisição HTTP. A

busca de dados na API ocorre pela *URL*: *http://compras.dados.gov.br/licitacoes/v1/licitacoes.-{formato}?{parametro1=valor1}*. Os valores entre chaves são os parâmetros que podem variar para realizar a busca de dados de acordo com condições específicas e no formato desejado, podendo compreender *XML, JSON, CSV e HTML* possibilitando uma visualização através do navegador e uma melhora na classificação desses conteúdos também nas ferramentas de busca automatizadas. Assim, a cada busca sobre a *API*, a *URL* de requisição é gerada com base na variação da entidade a ser utilizada no campo de pesquisa, atuando como um filtro predefinido. Como mencionado anteriormente, a aquisição dos registros para formar a base desta primeira base de dados foi sobre os códigos das Unidades Administrativas de Serviços Gerais.

4.2.2 Busca dos Publicações de Licitações

Durante a fase de busca das publicações, a *URL* de requisição, que foi criada a partir do parâmetro de filtro da entidade em questão, teve seu emprego efetuado com o objetivo de realizar uma requisição *HTTP*, resultando na obtenção dos dados em formato *CSV* por meio da técnica de *web scraping*. Foi possível proceder à extração das *URLs* das buscas retornadas.

4.2.3 Extração das Publicações de Licitações

Na sequência, foram realizadas requisições *HTTP* individuais para obter os dados em formato *CSV* das licitações correspondentes, utilizando a lista de *URLs* previamente identificadas. Com o emprego da técnica de *web scraping* e a biblioteca Pandas, procedeu-se à extração dos textos das entidades, armazenando-os de maneira organizada para análises subsequentes. A Figura 4.22 ilustra um exemplo de uma publicação do site de compras governamentais e suas respectivas publicações no portal do DOU. É possível notar, em um primeiro momento, a característica relacional dos dados presentes no site de compras, para que seja possível maior detalhamento em relação às entidades, de forma estruturada. Por este motivo, a primeira base de dados é formada por um arquivo *CSV* com as entidades nomeadas e seus respectivos valores.

4.3 Enriquecimento da Base de Publicações

Esta etapa de enriquecimento da base de publicações revelou-se uma etapa essencial para aprimorar a qualidade e relevância das informações contidas nela. O objetivo consistiu em utilizar entidades nomeadas, identificadas na etapa anterior, para realizar buscas por publicações de licitações específicas no banco de dados da ferramenta *Deep vacuity*. A abordagem adotada baseou-se em técnicas que empregam filtragem *SQL* e *fuzzy matching* com um limiar de 90% de similaridade, para faixa de corte. Em uma descrição breve, o *fuzzy matching* é uma técnica utilizada para comparar strings, levando em consideração a semelhança entre

Figura 4.22 – Publicação de uma licitação na API de Compras Governamentais (2023) e suas respectivas publicações do DOU.

elas, mesmo em casos de ortografia ou digitação ligeiramente diferentes. Essa abordagem é particularmente útil em situações em que se deseja encontrar correspondências aproximadas entre strings, sendo amplamente aplicada em tarefas como deduplicação de dados e correspondência de registros, conforme descrito por Jaro (1989). A medida de similaridade é calculada através de algoritmos que atribuem pontuações com base na proximidade das strings. Essas abordagens permitiram uma busca mais flexível e abrangente, levando em consideração variações e possíveis erros tipográficos nas entidades nomeadas. A fusão dessas técnicas proporcionou uma análise mais refinada e precisa, resultando em um conjunto de publicações mais elaborado.

Uma vez identificadas as publicações correlatas, o próximo passo foi a organização e armazenamento dessas informações. Os dados foram salvos em arquivos distintos, seguindo uma estrutura lógica que agrupa publicações semelhantes em diretórios correspondentes aos rótulos atribuídos. Essa abordagem não apenas facilita a gestão e busca futura, mas também permite manter a coesão entre as entidades nomeadas e as publicações associadas. É importante ressaltar que os resultados obtidos nessa fase foram integrados à base de dados em formato textual, mantendo a consistência e homogeneidade com a estrutura das publicações já existentes no DOU.

4.4 Validação das Publicações

Durante a etapa anterior, é possível que algumas publicações sem relação com a licitação desejada sejam incluídas, mesmo com o mapeamento eficiente das entidades. Essas publicações podem impactar negativamente o desempenho dos modelos a serem treinados, levando a rotulações incorretas ou duplicadas. Para mitigar esse problema, foi implementada esta etapa de validação das publicações após a busca no banco de dados da ferramenta *deep vacuity*. Essa etapa assegura que as publicações pertençam, de maneira única, à licitação em questão. O fluxograma desta etapa, representado na Figura 4.23, utiliza as entidades mapeadas como frequentes nas publicações, mas não utilizadas na busca, além de um hash md5 gerado na etapa anterior. Para cada entidade, verifica-se sua presença na publicação e

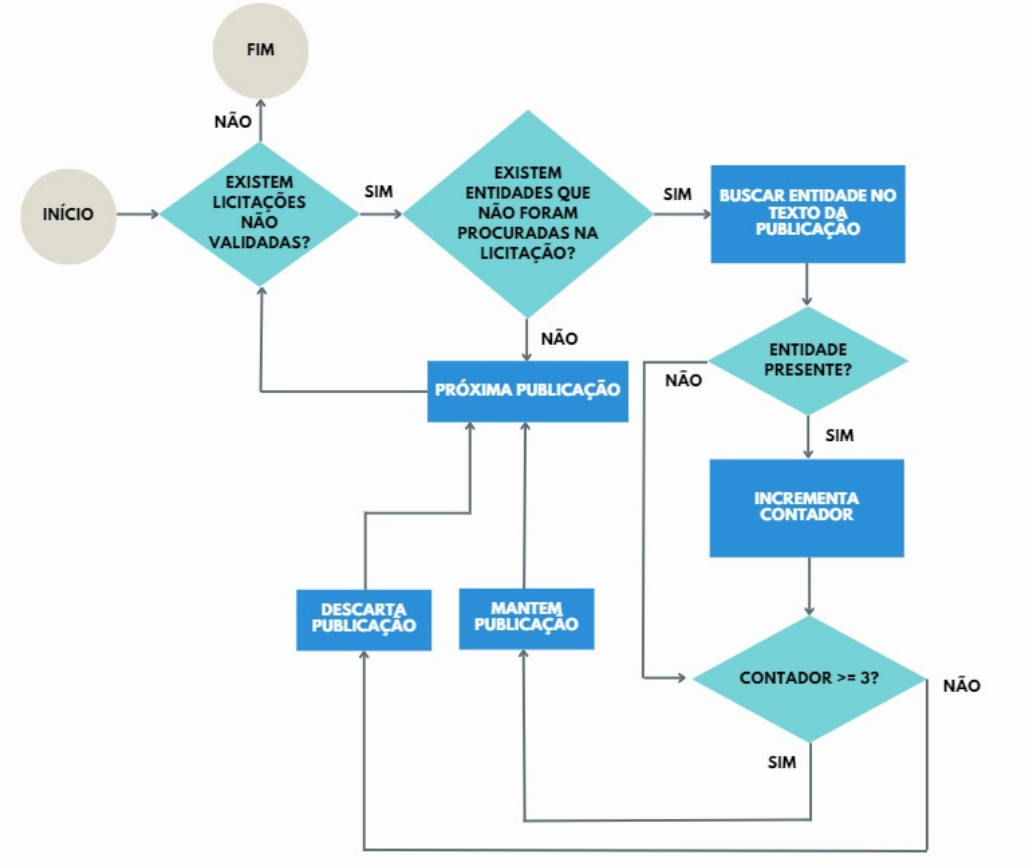

Figura 4.23 – Fluxograma de processos abordados na etapa de validação das publicações.

incrementa-se um contador, se positivo. Se pelo menos três entidades estiverem presentes, considera-se que a publicação pertence à licitação, caso contrário, ela é descartada. A escolha do valor mínimo de três entidades foi baseada em uma abordagem empírica durante o mapeamento, evitando problemas associados a valores menores ou maiores.

4.5 Treinamento

4.5.1 BERT

O modelo BERT, apresentado por Devlin et al. (2018), destacou-se como o precursor entre os modelos baseados em Transformers ao empregar uma abordagem que capacita o modelo a processar textos de forma integral, abrangendo tanto o contexto à esquerda quanto à direita, conforme representado na Figura 4.24. Essa estratégia envolve a utilização de um codificador de transformador bidirecional, treinado por meio de modelagem de linguagem mascarada. Essa inovação propiciou a conquista do estado-da-arte em tarefas desafiadoras de PLN, nas quais a consideração do contexto global do texto é fundamental.

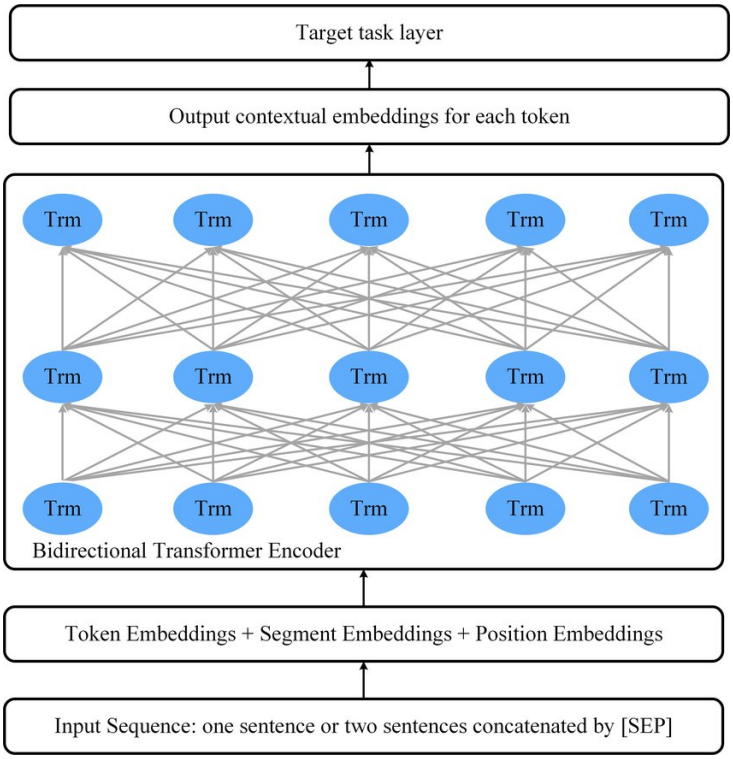

Figura 4.24 – Arquitetura do modelo BERT.

Fonte: Zhichang Zhang et al. (2019)

O primeiro modelo utilizado no trabalho é o BERT multilíngue, uma variante do BERT tradicional, que é um modelo baseado em Transformers pré-treinado em um amplo conjunto de dados em várias línguas, inclusive português, de maneira auto-supervisionada. Isso implica que o modelo foi pré-treinado apenas nos textos brutos, sem qualquer rotulagem humana, permitindo a utilização de uma quantidade considerável de dados públicos, segundo Hugging Face (2023a). Esse processo automático gera entradas e rótulos a partir desses textos. Especificamente, o modelo foi pré-treinado com dois objetivos principais:

- **Modelagem de linguagem mascarada (MLM):** Nesse processo, o modelo aleatoriamente mascara 15% das palavras na entrada de uma sentença e, em seguida, faz a

previsão das palavras mascaradas ao executar toda a sentença mascarada pelo modelo. Esse método permite ao modelo aprender uma representação bidirecional da sentença.

- **Previsão da próxima sentença (NSP):** O modelo concatena duas sentenças mascaradas como entradas, durante o pré-treinamento, e então precisa prever se as duas sentenças estavam sequenciais ou não.

Dessa forma, o modelo aprende uma representação interna das línguas no conjunto de treinamento, que pode ser utilizada para extrair características úteis em tarefas específicas. Ele é projetado para tarefas que utilizam sentenças completas para tomar decisões, como classificação de sequência, classificação de tokens ou resposta a perguntas.

4.5.2 DistilBERT

O modelo DistilBERT multilíngue é uma versão destilada do modelo BERT original. Assim como seu precursor, ele é baseado em Transformers pré-treinado em um amplo conjunto de dados multilíngues, também incluindo a língua portuguesa, de maneira auto-supervisionada. Além disso, sua arquitetura *cased* preserva as distinções entre maiúsculas e minúsculas, o que pode ser necessário em tarefas linguísticas sensíveis a casos, como afirma Hugging Face (2023b).

De acordo com Sanh et al. (2019), esta abordagem de destilação implica que o modelo foi compactado, mantendo um desempenho competitivo em tarefas *downstream*, ou seja, uma vez que o modelo tenha adquirido esse conhecimento geral, ele pode ser afinado ou ajustado para tarefas mais específicas enquanto reduz a complexidade e o tamanho do modelo em comparação com o BERT. O processo de destilação do modelo envolve a simplificação do BERT, mantendo as camadas essenciais e removendo redundâncias. Isso é alcançado através da utilização de uma técnica de aprendizado chamada destilação de conhecimento. Nesse processo, o modelo BERT atua como um professor, transmitindo seu conhecimento para o modelo menor, DistilBERT. O objetivo é que o modelo destilado possa capturar a essência das representações aprendidas pelo modelo maior, mas com uma estrutura mais compacta.

Ao contrário do BERT, o DistilBERT simplifica o processo de treinamento e a arquitetura do modelo, tornando-o mais leve e eficiente. Essa característica o torna especialmente útil em cenários em que recursos computacionais são limitados.

5 Resultados

Neste capítulo, serão discutidos os resultados do estudo em quatro seções distintas. A primeira abordará o mapeamento das entidades de licitações, a segunda dará ênfase nas etapas de busca, validação e construção da base de dados rotulada, a terceira apresentará os resultados do treinamento e uma comparação entre os dois modelos de Transformer, enquanto a quarta fornece uma visão geral das etapas anteriores e discute as melhores abordagens para cada contexto específico.

O trabalho foi desenvolvido em um servidor de alto desempenho fornecido pelo Departamento de Ciência da Computação (CIC), capaz de realizar os treinamentos necessários devido à sua capacidade computacional. Suas especificações podem ser vistas na Tabela 5.4.

Componente	**Especificação**
Modelo do Servidor	GPX XS3-11S1-2GPU
Configuração RAID	Intel C621 (AHCI) [SATA 6, 6 portas] RAID 1 - 2 x 8.0TB SATA 6.0Gb/s 7200RPM - 3.5- Ultrastar DC HC320 (512e)
M.2 Sockets	M Key [NVMe, 2 portas] Sem RAID (*OS) - 1 x 1.0TB Toshiba XG6 M.2 PCIe 3.1 x4 NVMe Solid State Drive
Chipset	Intel C621
Tipo de Servidor	1U GPU Server
Conectividade	Dual 10-Gigabit Ethernet
Fonte de Alimentação	1400W Single Power Supply
Processador	Intel Xeon Gold 5220R, 24-Core 2.2GHz, 35.75MB Cache (150W)
Memória RAM	6 x 64GB PC4-23400 2933MHz DDR4 ECC RDIMM
Armazenamento SSD	1.0TB Toshiba XG6 M.2 PCIe 3.1 x4 NVMe Solid State Drive
Armazenamento HDD	2 x 8.0TB SATA 6.0Gb/s 7200RPM - 3.5- Ultrastar DC HC320 (512e)
GPUs	2 x NVIDIA Tesla V100S GPU Computing Accelerator - 32GB HBM2 - PCIe 3.0 x16 - Passive Cooling
Outros Componentes	AC Power Cord (North America), C13, NEMA 5-15P, 2.1m CAB-AC

Tabela 5.4 – Especificações do Servidor GPX XS3-11S1-2GPU.

5.1 Mapeamento das Entidades

Os resultados da análise das entidades de licitações estão disponíveis na Tabela 5.5. Como discutido no capítulo anterior, foram identificadas as características de unicidade e frequência de cada entidade durante a pesquisa por publicações. Para uma melhor compreensão, os resultados foram categorizados em três níveis: baixo, médio e alto. O nível baixo indica escassa ou nenhuma unicidade ou frequência da entidade, ao passo que o nível alto indica uma alta presença de unicidade ou frequência.

Entidade	Unicidade	Frequência
codigo do item no catalogo	Baixo	Média
data abertura proposta	Baixo	Alta
data entrega edital	Baixo	Alta
data entrega proposta	Baixo	Alta
data publicação	Baixo	Alta
endereço entrega edital	Baixo	Alta
função responsável	Baixo	Baixa
identificador	Alta	Alta
informações gerais	Baixa	Alta
modalidade	Baixa	Média
nome responsável	Baixa	Baixa
número aviso	Médio	Médio
número item licitação	Baixa	Baixa
número itens	Baixa	Alta
número processo	Alta	Alta
objeto	Alta	Médio
situação aviso	Baixa	Baixa
tipo pregão	Baixa	Médio
tipo recurso	Baixa	Baixa
uasg	Baixa	Alta

Tabela 5.5 – Resultado do mapeamento das entidades.

Inicialmente, para obter publicações relacionadas a licitações no site de compras governamentais, priorizou-se entidades com baixa unicidade e alta frequência. Foram selecionadas apenas as entidades com características de baixa unicidade e alta frequência para aquisição do maior número de publicações da API de Compras Governamentais (2023), tais como:

- Código da uasg.
- Data de publicação.

No caso das pesquisas em Imprensa Nacional (2023) e no banco de dados de publicações, a unicidade foi essencial para a eficácia da busca. A unicidade média ou alta

foi considerada mecessária, uma vez que uma unicidade baixa resultaria em um grande número de publicações não relacionadas à licitação, tornando a busca demorada e ineficiente. Portanto, foram escolhidas apenas entidades com unicidade média ou alta como:

- Identificador da licitação.
- Número do aviso da licitação.
- Número do processo da licitação.
- Objeto da licitação.

Na validação das publicações, a frequência tornou-se o principal critério, assegurando a similaridade entre o objeto desejado e o buscado. Somente entidades com frequência média ou alta foram escolhidas para a validação, excluindo aquelas utilizadas na busca. Nesta etapa, foram escolhidas as entidades:

- Data de publicação.
- Data de abertura proposta.
- Data de entrega edital.
- Data de entrega da proposta.
- Data da publicação.
- Endereço de entrega do edital.
- Informações gerais.
- Modalidade da licitação.
- Número de itens da licitação.
- Código do item no catálogo.

Entidades como Tipo de recurso, Número do item da licitação, Situação do aviso, Nome do responsável e Função do responsável foram excluídas nas etapas de busca e validação devido às baixas propriedades de unicidade e frequência. Apesar de não serem adequadas para a coleta de publicações, essas entidades ainda detêm informações relevantes sobre as licitações e compõem a base de dados rotulada.

5.2 Base de Publicações

Após as etapas de busca, validação e aquisição de publicações, foi constituída uma base de 191.442 publicações rotuladas, pertencentes a 15.552 classes distintas. A Figura 5.25 exibe três histogramas com os resultados quantitativos da base de dados obtida.

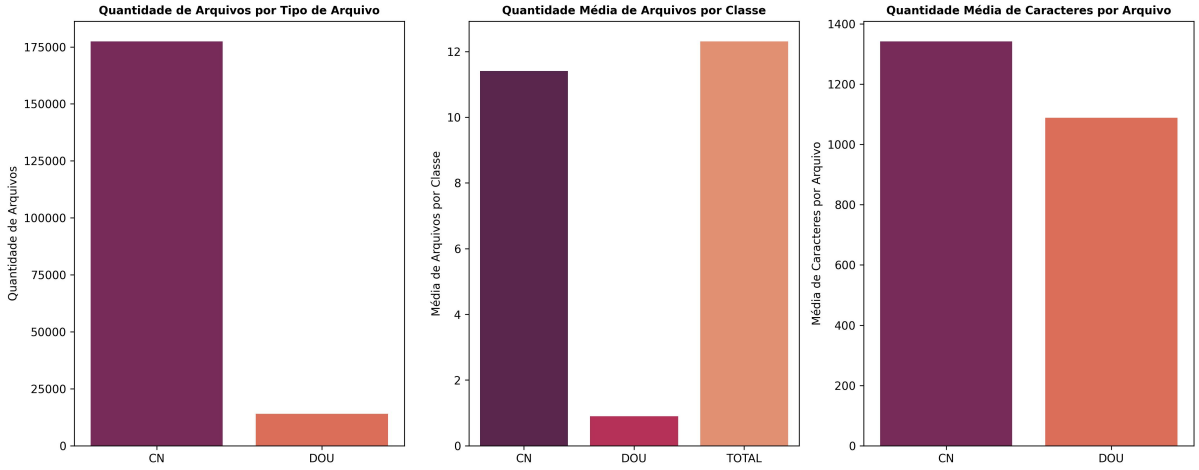

Figura 5.25 – Histogramas de publicações totais, médias e de quantidade de caracteres.

O primeiro histograma apresenta a quantidade de publicações segmentadas por tipo de origem: comprasnet ou banco de dados do DOU. Foram adquiridas 177.466 publicações do comprasnet, em contraste com 13.976 do diário oficial. Essa diferença notável ocorre porque as publicações do DOU não especificam discricionariamente os itens licitados, apenas seu número total, ao contrário das publicações do comprasnet, gerando uma relação de desmembramento e correspondência múltipla, conforme será descrito no próximo gráfico.

O segundo histograma descreve a quantidade média de arquivos por classe de treinamento, ou seja, seus rótulos. Obteve-se uma média de aproximadamente 12 publicações associadas a cada classe, com representatividade de 11 publicações do comprasnet para 1 do DOU. Esses números oferecem uma visão mais genérica da base gerada, considerando casos em que rótulos foram formados por mais de uma publicação do DOU ou exclusivamente do comprasnet, devido à falta de paridades no banco de publicações durante a etapa de aquisição.

O terceiro histograma representa a quantidade média de caracteres por tipo de origem das publicações em uma amostra de classes distintas. Identificou-se que o texto nas publicações do DOU possui, em média, 1.088 caracteres, enquanto os textos gerados a partir de publicações do comprasnet têm, em média, 1.342 caracteres.

Como especificado, a base de dados possui características peculiares relacionadas às publicações pertencentes a cada classe, que podem ser compostas exclusivamente por publicações do diário, do comprasnet ou por ambas. Essa estratégia foi adotada para reforçar o contexto de aplicação real, considerando que nem sempre há uma relação clara entre

as publicações de contextos distintos, e o modelo deve ser capaz de atribuir classificações mesmo nessas situações.

5.3 Treinamento

Nesta seção serão apresentados, de forma sucinta, os resultados do treinamento de cada um dos dois modelos treinados neste trabalho: BERT e DistilBERT. Em seguida, será realizado um comparativo entre eles.

5.3.1 Modelo BERT multilíngue

Conforme mencionado em seções anteriores, o modelo BERT multilíngue é uma implementação do BERT que oferece suporte a diversas línguas, incluindo o português. No caso específico do pré-treinamento com dados na língua portuguesa, este modelo passou por um processo de aprendizado prévio em grandes volumes de textos em português, visando capturar padrões linguísticos, semânticos e contextuais específicos desse idioma. Durante essa fase, o modelo aprende a representar de maneira eficaz as relações entre palavras e a compreender nuances linguísticas presentes na nossa língua.

Os resultados do treinamento desse modelo estão representados em dois gráficos na Figura 5.26. O primeiro gráfico ilustra a evolução das métricas - acurácia, sensibilidade, precisão e F1-Score - ao longo das épocas de treinamento. O segundo gráfico exibe as curvas de perdas relacionadas ao treinamento e ao próprio modelo. Observa-se que, ao término do treinamento, o modelo alcançou um bom desempenho na tarefa de classificação multiclasse em dados de publicações de licitação, conforme detalhado na Tabela 5.6. Após um treinamento abrangente, as métricas mais destacadas foram 0,952100 para acurácia e sensibilidade, 0,948564 para precisão e 0,948380 para o F1-Score. Esses resultados indicam um alto nível de eficácia do modelo na classificação de dados, demonstrando consistência e robustez em relação às métricas avaliadas.

No contexto das métricas observadas no gráfico, destaca-se que, nas primeiras 20 épocas de treinamento, o modelo exibe uma discreta diferença numérica, na ordem centesimal, que se torna evidente ao comparar as curvas de acurácia com as de precisão e F1-Score. Essa disparidade pode ser atribuída à detecção de Falsos Positivos (FP) durante essa fase do treinamento, indicando que algumas publicações foram erroneamente classificadas. Esse fenômeno impacta diretamente a precisão, uma vez que, matematicamente, a introdução de falsos positivos reduz esse valor, o que, por sua vez, se reflete no F1-Score.

Outro ponto relevante a ser observado é que tanto a acurácia quanto a sensibilidade mantiveram valores iguais ao longo das épocas. Essa constância é observada quando o número de Verdadeiros Negativos (VN) e Falsos Negativos (FN) é nulo. Essa condição específica surge quando o modelo realiza previsões corretas para todas as instâncias positivas.

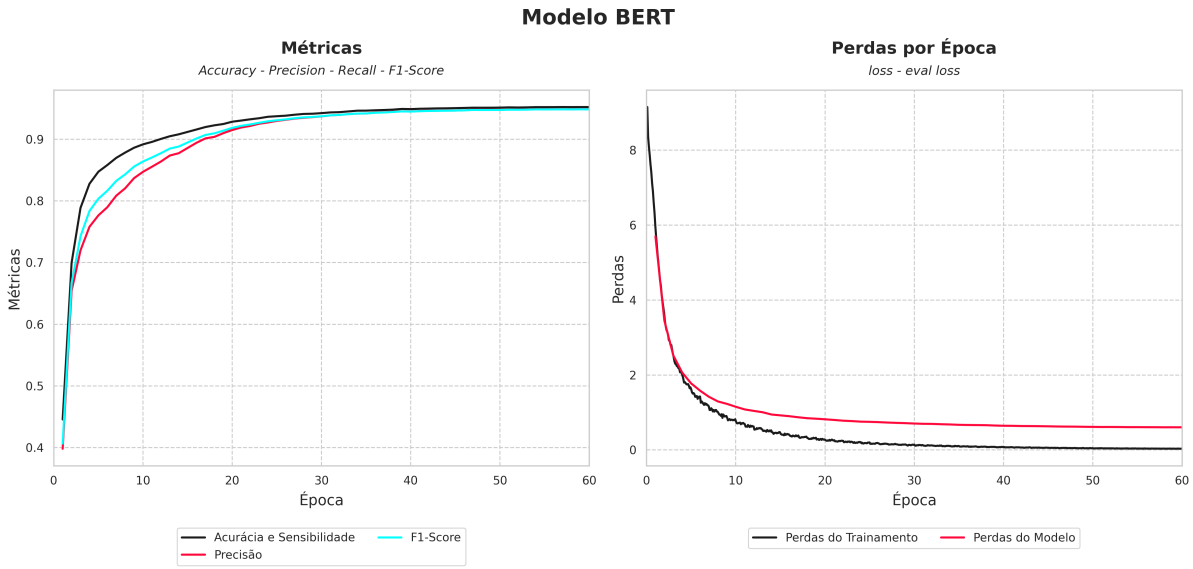

Figura 5.26 – Evolução das métricas ao longo das épocas para o modelo BERT.

5.3.2 Modelo DistilBERT multilíngue

Como o modelo DistilBERT multilíngue representa uma versão compacta e eficiente do modelo anterior, ele possui as mesmas características em relação e exposição a diversas línguas, inclusive o português. Ele passa por um processo de destilação que visa reduzir a complexidade do modelo original mantendo um desempenho próximo. Esse procedimento envolve a transferência de conhecimento do modelo mais complexo para o mais compacto. A destilação resulta em uma redução no número de parâmetros e, consequentemente, em uma menor carga computacional durante a inferência. Embora o DistilBERT possa sacrificar um pouco de desempenho em comparação com o modelo original, essa perda é compensada pela vantagem da eficiência computacional.

Os resultados do treinamento desse modelo estão representados em dois gráficos na Figura 5.27. O primeiro gráfico ilustra a evolução das métricas - acurácia, sensibilidade, precisão e F1-Score - ao longo das épocas de treinamento. O segundo gráfico exibe as curvas de perdas relacionadas ao treinamento e ao próprio modelo. Observa-se que, ao término do treinamento, o modelo também alcançou um bom desempenho na tarefa de classificação multiclasse em dados de publicações de licitação, conforme detalhado na Tabela 5.6. Após um treinamento abrangente, as métricas mais destacadas foram 0,946511 para acurácia e sensibilidade, 0,945880 para precisão e 0,943783 para o F1-Score. Esses resultados indicam um alto nível de eficácia do modelo na classificação de dados, demonstrando consistência e robustez em relação às métricas avaliadas.

As curvas de métricas apresentadas para este modelo são bem semelhantes e próximas, caracterizando uma sutil diferença numérica, na ordem milesimal, que não se notam apenas analisando o gráfico de forma visual. Aqui também foi observado o fato de que tanto a

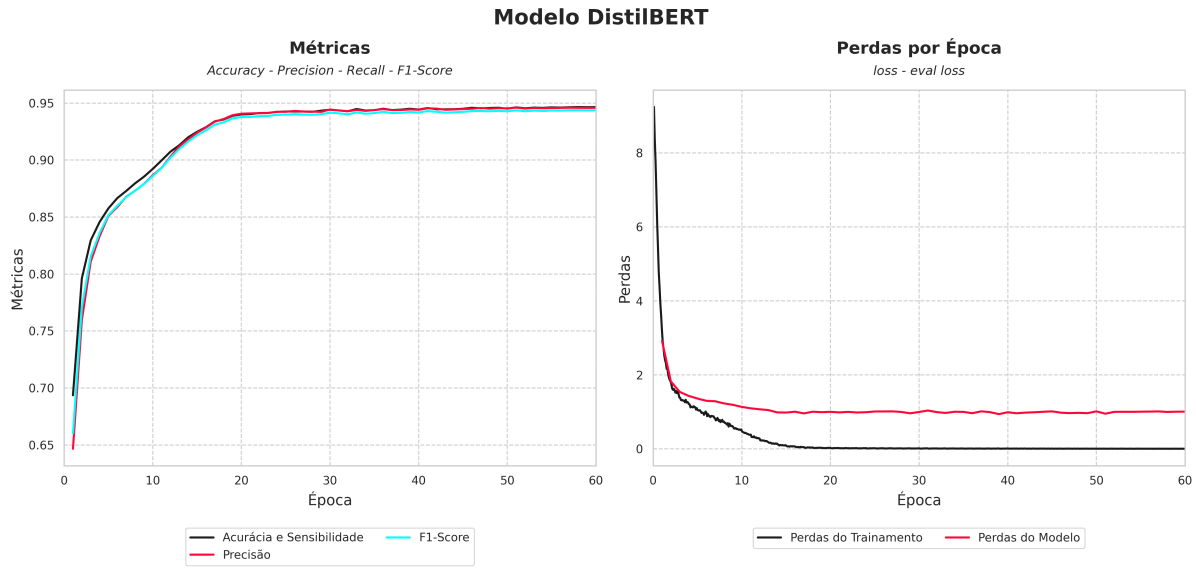

Figura 5.27 – Evolução das métricas ao longo das épocas para o modelo DistilBERT.

acurácia quanto a sensibilidade mantiveram valores iguais ao longo das épocas, seguindo o mesmo padrão.

5.3.3 Comparativo

Como pode ser analisado na Tabela 5.27, a diferença numérica das métricas entre ambos os modelos é mínima, da ordem milesimal.

Modelo de Transformer	Acurácia	Sensibilidade	Precisão	F1-Score
BERT	0,952100	0,952100	0,948564	0,948380
DistilBERT	0,946511	0,946511	0,945880	0,943783

Tabela 5.6 – Tabela de melhores métricas de treinamento para cada modelo.

As principais diferenças observadas entre os modelos, presentes nos gráficos de métricas da Figura 5.28, residem em seus valores extremos. Ambos os modelos, BERT e DistilBERT, obtiveram F1-Scores muito parecidos, registrando valores de 0,948 e 0,943, respectivamente. Uma diferença notável durante os treinamentos foi que o DistilBERT mostrou um desempenho superior nas fases iniciais, ao passo que o BERT se destacou nas etapas finais. Embora não tenha sido o foco central desta pesquisa, vale ressaltar que o modelo destilado demonstrou uma execução mais rápida e menor custo computacional, características inerentes a essa abordagem, embora sem garantias absolutas.

A mesma ideia de desempenho também pode ser avaliada na Figura 5.29, que demonstra as perdas do treinamento e do modelo. Nas épocas iniciais o DistilBERT apresenta melhor desempenho, menos perdas. Já nas épocas finais essa perspectiva é invertida e o modelo BERT tem melhor performace, menos perdas.

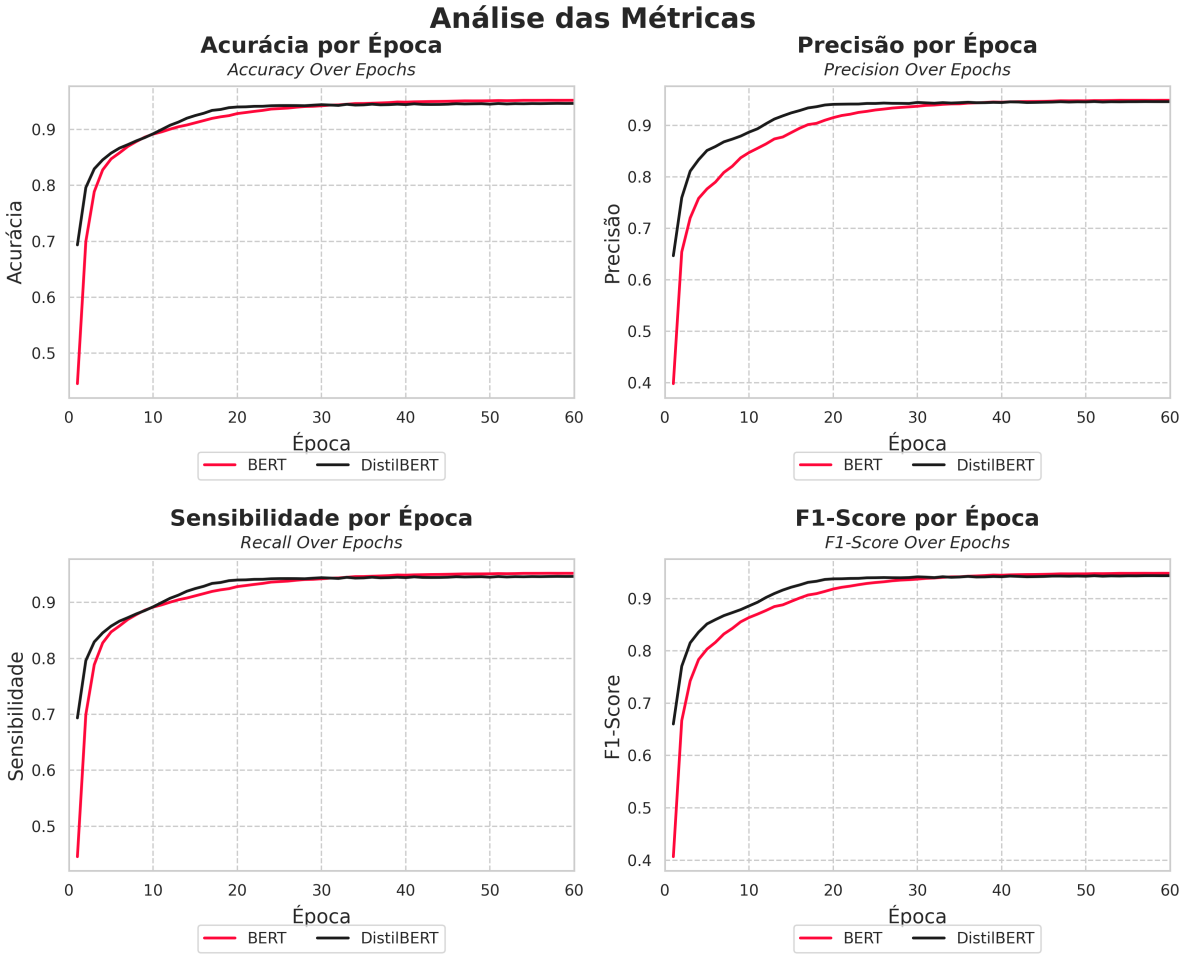

Figura 5.28 – Comparativo entre a evolução das métricas ao longo das épocas.

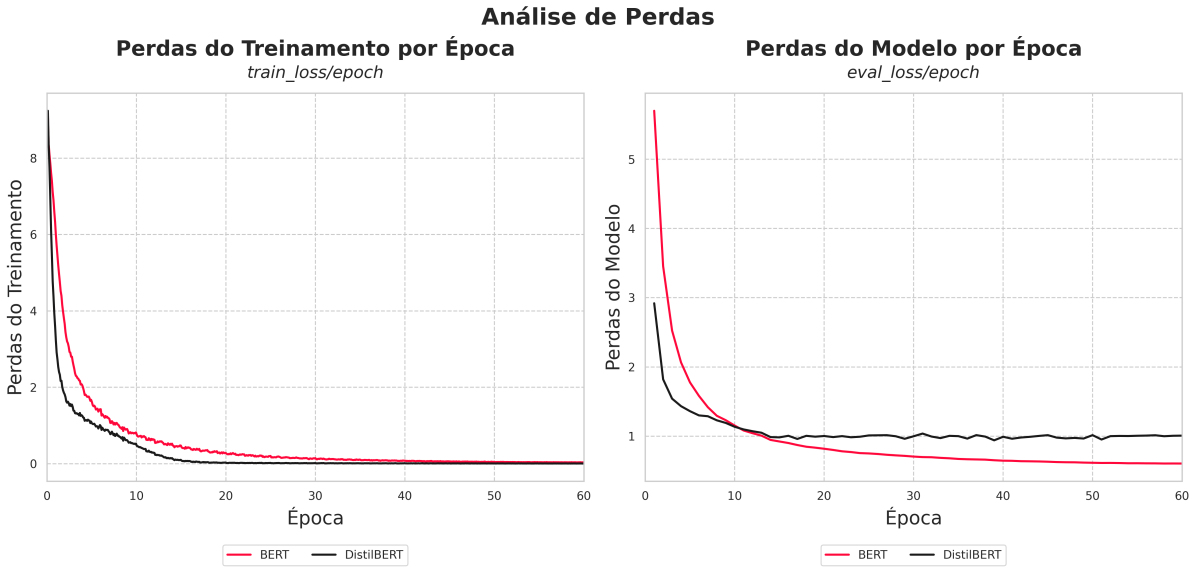

Figura 5.29 – Comparativo entre a evolução das perdas ao longo das épocas.

Ainda, como ambos os modelos foram treinados sobre o mesmo conjunto de dados

utilizando os mesmo parâmetros, a Figura 5.30, primeiro gráfico, ilustra o decaimento linear da taxa de aprendizado para ambos os modelos de maneira constante ao longo das épocas. Essa característica é responsável por favorecer a convergência e o refinamento do modelo. O segundo gráfico por sua vez, também com característica linear, demonstra o número de passos que definem uma época de treinamento, 5.584 passos. Esse parâmetro é função do conjunto de dados de treinamento e do tamanho do lote escolhido.

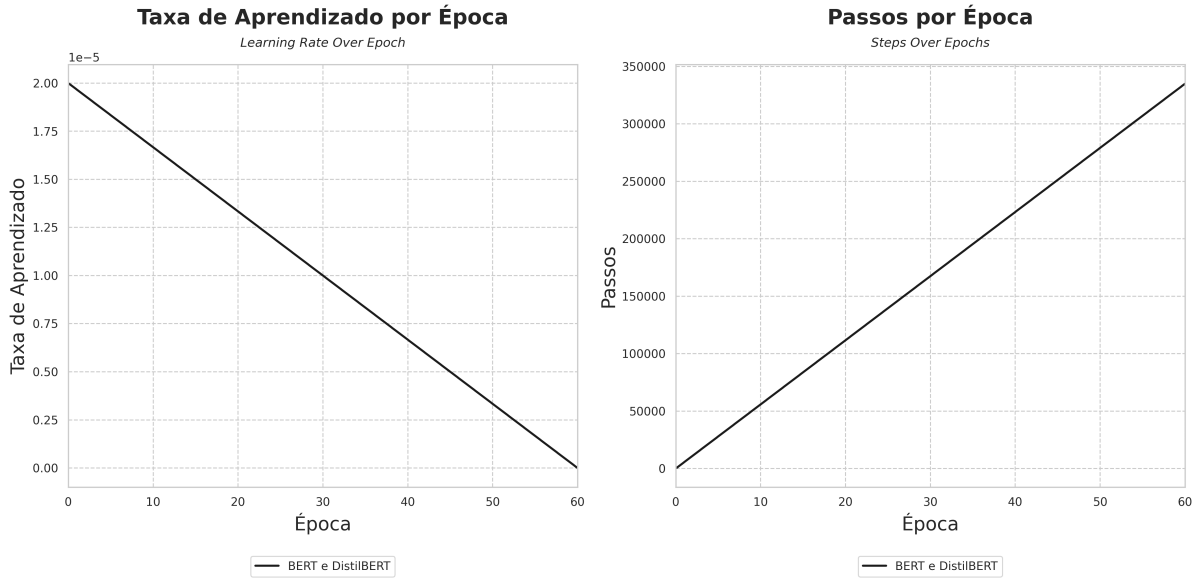

Figura 5.30 – Curvas de taxa de aprendizado e numero de passos em função das épocas.

5.4 Discussão dos Resultados

Os resultados deste estudo revelam a eficácia da abordagem adotada na construção da base de dados, apesar do claro desbalanceamento entre essas classes. Embora a disparidade na representação das classes pudesse ser um desafio, os resultados satisfatórios obtidos na fase de treinamento dos Transformers indicam que a técnica utilizada para ponderar as métricas em função desse desequilíbrio foi bem-sucedida. Isso sugere que a base de dados construída proporcionou um ambiente propício para o aprendizado dos modelos, resultando em performances consistentes.

No que diz respeito à comparação entre os modelos BERT e DistilBERT, ambos alcançaram F1-Scores muito próximos. A distinção mais notável entre os dois modelos foi a variação de desempenho ao longo das épocas de treinamento, com o DistilBERT apresentando superioridade nas fases iniciais e o BERT destacando-se nas etapas finais.

Em um contexto mais amplo, os resultados sugerem que o modelo BERT tende a oferecer um desempenho superior para o propósito geral, manifestando-se com um F1-Score ligeiramente mais elevado. Contudo, é necessário considerar que, em situações em que a rapidez na obtenção de resultados é prioritária e um treinamento mais eficiente é

desejado com um menor número de épocas, o modelo DistilBERT tende a se destacar. Suas características intrínsecas, como execução mais rápida e menor custo computacional, tornam-no uma escolha viável para cenários que demandam eficiência temporal sem comprometer significativamente a qualidade do modelo. No entanto, é importante ressaltar que essas características, embora promissoras, não estão isentas de incertezas absolutas. A escolha entre BERT e DistilBERT, portanto, dependerá das prioridades específicas e das exigências do contexto de aplicação.

A aplicação de Transformers para agrupar informações de licitações provenientes de fontes heterogêneas, representa uma abordagem inovadora e eficaz para a detecção de fraudes em atos públicos. Dada a volumosa quantidade de dados diariamente inseridos nos diários públicos, a capacidade desses modelos em compreender e relacionar essas informações possibilita a criação de grupos coerentes e abrangentes. Esse agrupamento facilita a análise e identificação de padrões suspeitos, permitindo uma avaliação mais rápida e precisa das licitações. A sinergia entre as fontes de dados contribui significativamente para revelar relações ocultas e comportamentos irregulares, potencializando a efetividade das investigações anticorrupção e fortalecendo a transparência nos processos governamentais. A utilização dessas tecnologias não apenas otimiza o trabalho dos órgãos responsáveis, mas também promove um ambiente mais íntegro e ético na condução dos negócios públicos.

6 Conclusões

Este trabalho teve como objetivo treinar modelos de Transformers para a classificação de entidades nomeadas, relativas à publicações de processos licitatórios públicos de fontes heterogêneas. Para atingir essa meta, foi essencial criar uma base de publicações rotuladas para permitir o aprendizado dos modelos. A construção eficiente e eficaz dessa base de dados envolve algumas etapas, visando proporcionar um conjunto de dados de alta qualidade para o treinamento dos modelos.

Na etapa inicial, destacou-se a necessidade de compreender as entidades presentes nos textos das publicações. Ao mapear essas entidades, tornou-se viável determinar quais eram mais apropriadas para as fases de busca e validação das publicações, além de examinar as diversas formas de ocorrência de cada entidade nos textos. Foi observado que as entidades com unicidade, isto é, aquelas que servem como identificadores eficazes de uma licitação, demonstraram ser as mais eficientes na localização de publicações do DOU. Por outro lado, concluiu-se que as entidades que surgem com maior frequência nas publicações são mais adequadas para fins de validação.

Nas etapas seguintes, abordou-se a técnica empregada na busca e validação das publicações. Essas fases, aliadas à rotulação, resultaram em uma base contendo 191.442 publicações distribuídas em 15.552 classes distintas. Destacou-se uma variação significativa na quantidade de publicações relacionadas a cada classe, originando-se de diferentes fontes. As classes foram formadas de três maneiras distintas, podendo cada uma ser composta exclusivamente com publicações do DOU, exclusivamente com publicações da API de Compras Governamentais ou com ambas.

Apesar do desequilíbrio nas classes, a técnica empregada para ajustar as métricas com base nessa disparidade de amostras provou ser eficaz, resultando em um desempenho bastante satisfatório após o treinamento dos modelos. Ambos os modelos, BERT e DistilBERT, alcançaram F1-Scores bastante semelhantes, com valores de 0,948 e 0,943, respectivamente. Uma distinção notável entre os treinamentos reside no fato de que o modelo DistilBERT demonstrou um desempenho superior nas épocas iniciais, enquanto o modelo BERT se destacou nas épocas finais. Embora não tenha sido o foco deste estudo, observou-se que o modelo destilado apresentou menor tempo de execução e custo computacional, características inerentes a essa abordagem, embora sem garantias absolutas.

A aplicação de técnicas baseadas em Transformers simplifica a detecção de fraudes em licitações públicas, ao agrupar eficientemente informações de fontes heterogêneas. Isso agiliza a identificação de padrões suspeitos, fortalecendo investigações anticorrupção e promovendo transparência nos processos governamentais.

Para trabalhos futuros, a intenção é equilibrar as classes na base de publicações rotuladas e realizar os treinamentos em uma quantidade maior de épocas, visando aprimorar o aprendizado dos modelos. Outra possibilidade de pesquisa está relacionada à exploração de diferentes modelos de Transformer, como RoBERTa, ALBERT, XLNet, entre outros, durante a fase de treinamento. Embora este estudo tenha se concentrado nas publicações de licitações, ele serve como ponto de partida para pesquisas mais abrangentes que incluirão uma variedade maior de instrumentos públicos, como convênios e contratos, além de considerar outros tipos de entidades relevantes. Além disso, há a intenção de aprimorar o modelo para compreender as relações entre licitações, convênios e contratos.

Referências

AGIRRE, E.; EDMONDS, P.; WORDEN, R. Word sense disambiguation: a survey. **Computational Linguistics**, MIT Press, v. 33, n. 4, p. 371–416, 2007. Citado na p. 22.

ALBANAZ, J. O. L. **Reconhecimento de Entidades Nomeadas em resultados de licitações publicados em Diários Oficiais**. 2020. Monografia (Especialização) – Universidade Federal do Paraná, Curitiba. Citado na p. 13.

ANDREAS C. MÜLLER, S. G. **Introduction to Machine Learning with Python: A Guide for Data Scientists**. 1. ed.: O'Reilly Media, 2016. Citado nas pp. 33–35.

API DE COMPRAS GOVERNAMENTAIS. **Dicionário de Dados - Licitações**. Acesso em: 31 ago. 2023. 2023. Citado nas pp. 14, 40, 42, 44, 47, 52.

BA, J. L.; KIROS, J. R.; HINTON, G. E. **Layer Normalization**. 2016. arXiv: 1607.06450 [stat.ML]. Citado na p. 31.

BENABDELLAH, A. C.; BENGHABRIT, A.; BOUHADDOU, I. A survey of clustering algorithms for an industrial context. **Procedia Computer Science**, v. 148, p. 291–302, 2019. THE SECOND INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTING IN DATA SCIENCES, ICDS2018. Citado na p. 19.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural Language Processing with Python**. O'Reilly Media, 2009. Citado na p. 19.

BRASIL. **Lei n. 14.133, de 1 de abril de 2021**. 2021. https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14133.htm. Acesso em: 29 ago. 2023. Citado na p. 13.

BRASIL. **Lei n. 8.666, de 21 de junho de 1993**. 1993. http://www.planalto.gov.br/ccivil_03/leis/l8666cons.htm. Acesso em: 29 ago. 2023. Citado na p. 13.

BRASIL. **Medida Provisória Nº 961, de 6 de maio de 2020**. 2020. https://www.in.gov.br/web/dou/-/medida-provisoria-n-961-de-6-de-maio-de-2020-255615815. Acesso em: 2 ago. 2023. Citado na p. 12.

BRITO, H. S. d.; COSTA, A. C. O. d. Corrupção em tempos da covid-19: o papel do Controle Externo nos desafios provocados pelo atual cenário pandêmico. **Revista Técnica dos Tribunais de Contas**, 2021. Citado na p. 12.

BROWN, T. B.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J.; DHARIWAL, P.; NEELAKANTAN, A.; SHYAM, P.; SASTRY, G.; ASKELL, A.; AGARWAL, S.; HERBERT-VOSS, A.; KRUEGER, G.; HENIGHAN, T.; CHILD, R.; RAMESH, A.; ZIEGLER, D. M.; WU, J.; WINTER, C.; HESSE, C.; CHEN, M.; SIGLER, E.; LITWIN, M.; GRAY, S.; CHESS, B.; CLARK, J.; BERNER, C.; MCCANDLISH, S.; RADFORD, A.; SUTSKEVER, I.;

AMODEI, D. **Language Models are Few-Shot Learners**. 2020. arXiv: 2005.14165 [cs.CL]. Citado na p. 39.

BÜYÜKÖZ, B.; HÜRRIYETOĞLU, A.; ÖZGÜR, A. Analyzing ELMo and DistilBERT on Socio-political News Classification. English. In: HÜRRIYETOĞLU, A.; YÖRÜK, E.; ZAVARELLA, V.; TANEV, H. (Ed.). **Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020**. Marseille, France: European Language Resources Association (ELRA), mai. 2020. P. 9–18. ISBN 979-10-95546-50-4. Disponível em: <https://aclanthology.org/2020.aespen-1.4>. Citado na p. 39.

CANDIDO, G. **Aprendizado supervisionado x não supervisionado**. Acesso em 05 dez. 2023. 2023. Disponível em: <https://medium.com/analytics-vidhya/natural-language-processing-nlp-based-chatbots-7b2436428256>. Citado na p. 18.

CGU. **CGU monitora aplicação dos recursos federais repassados a estados e municípios**. 2023. https://www.gov.br/cgu/pt-br/coronavirus/cgu-monitora-aplicacao-dos-recursos-federais-repassados-a-estados-e-municipios. Acesso em: 2 ago. 2023. Citado na p. 12.

CHEN, J.; LIU, X. Anomaly Detection in Public Procurement Data Using Statistical Methods. **European Journal on Criminal Policy and Research**, v. 25, n. 3, p. 333–349, 2019. Citado na p. 12.

COMPRANET. **Catalogo de Unidades Administrativas de Serviços Gerais.** Acesso em: 31 ago. 2023. 2023. Citado na p. 45.

DANIEL JURAFSKY, J. H. M. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. 3. ed., 2023. Citado nas pp. 25–28, 30–33, 35, 37.

DEVLIN, J.; CHANG, M.; LEE, K.; TOUTANOVA, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. **CoRR**, abs/1810.04805, 2018. arXiv: 1810.04805. Disponível em: <http://arxiv.org/abs/1810.04805>. Citado nas pp. 37, 49.

EINORYTĖ, A. **What is speech recognition, and how does it work?** Disponível em: <https://nordvpn.com/pt-br/blog/what-is-speech-recognition/>. Acesso em: 3 dez. 2023. Citado nas pp. 20, 21.

FACELI, K. **Inteligência artificial : uma abordagem de aprendizado de máquina**. Grupo Gen - LTC, 2011. P. 4–7. Citado na p. 17.

FORTINI, C.; MOTTA, F. Corrupção nas licitações e contratações públicas: sinais de alerta segundo a Transparência Internacional. **Revista de Direito Administrativo e Constitucional**, n. 64, p. 93–113, 2016. Citado na p. 12.

GIENAPP, L.; KIRCHEIS, W.; SIEVERS, B. et al. A large dataset of scientific text reuse in Open-Access publications. **Scientific Data**, v. 10, p. 58, 2023. DOI: 10.1038/s41597-022-01908-z. Citado na p. 24.

GOWRISANKAR, B.; THING, V. L. L. **An adversarial attack approach for eXplainable AI evaluation on deepfake detection models**. 2023. arXiv: 2312.06627 [cs.CV]. Citado na p. 39.

HARSHA, A. **Text Classification with BERT**. Acessado em 11 de dezembro de 2023. 2023. Disponível em: <https://www.shiksha.com/online-courses/articles/text-classification-with-bert/>. Citado na p. 24.

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. 31.

HOWARD, J.; RUDER, S. Universal language model fine-tuning for text classification. **arXiv preprint arXiv:1801.06146**, 2018. Citado na p. 37.

HUGGING FACE. bert-base-multilingual-cased. **Hugging Face Model Hub**, 2023a. Acesso em: 20 nov. 2023. Disponível em: <https://huggingface.co/bert-base-multilingual-cased>. Citado na p. 49.

HUGGING FACE. distilbert-base-multilingual-cased. **Hugging Face Model Hub**, 2023b. Acesso em: 21 nov. 2023. Disponível em: <https://huggingface.co/distilbert-base-multilingual-cased>. Citado na p. 50.

IBM. **What is natural language processing (NLP)?** Disponível em: <https://www.ibm.com/topics/natural-language-processing>. Acesso em: 30 ago. 2023. Citado nas pp. 19, 20.

IMPRENSA NACIONAL. **Diário Oficial da União: Pesquisa**. Acesso em: 31 ago. 2023. 2023. Citado nas pp. 14, 43, 44, 52.

JARO, M. A. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. **Journal of the American Statistical Association**, Taylor & Francis, v. 84, n. 406, p. 414–420, 1989. DOI: 10.1080/01621459.1989.10478785. Citado na p. 47.

JAYANTHI, K.; MAHESH, C. A Study on machine learning methods and applications in genetics and genomics. **International Journal of Engineering and Technology(UAE)**, v. 7, p. 201–204, fev. 2018. DOI: 10.14419/ijet.v7i1.7.10653. Citado nas pp. 18, 19.

JÚNIOR, D. G.; FILHO, G. S.; CABRAL, L. Classificação de fraudes em licitações públicas através do agrupamento de empresas em conluios. In: ANAIS do XI Workshop de Computação Aplicada em Governo Eletrônico. João Pessoa/PB: SBC, 2023. P. 13–24. DOI: 10.5753/wcge.2023.229519. Disponível em: <https://sol.sbc.org.br/index.php/wcge/article/view/24861>. Citado na p. 12.

KELLEHER, J. D.; MACNAMEE, B.; D'ARCY, A. **Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies**. Cambridge, MA: MIT Press, 2015. Citado na p. 17.

KHANAM, M.; MAHBOOB, T.; IMTIAZ, W.; GHAFOOR, H.; SEHAR, R. A survey on unsupervised machine learning algorithms for automation, classification and maintenance. **International Journal of Computer Applications**, v. 119, p. 34–39, jun. 2015. Citado na p. 19.

LEWIS TUNSTALL LEANDRO VON WERRA, T. W. **Natural Language Processing with Transformers**. 1. ed.: O'Reilly Media, Inc., 2021. Citado na p. 33.

LIANG, B.; LI, H.; SU, M.; BIAN, P.; LI, X.; SHI, W. Deep Text Classification Can be Fooled. In: PROCEEDINGS of the Twenty-Seventh International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, jul. 2018. (IJCAI-2018). DOI: 10.24963/ijcai.2018/585. Disponível em: <http://dx.doi.org/10.24963/ijcai.2018/585>. Citado na p. 39.

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.findingsemnlp.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. In. DOI: https://repositorio.unb.br/handle/10482/42026. Citado nas pp. 13, 40.

LIMSOPATHAM, N. Effectively Leveraging BERT for Legal Document Classification. In: ALETRAS, N.; ANDROUTSOPOULOS, I.; BARRETT, L.; GOANTA, C.; PREOTIUC-PIETRO, D. (Ed.). **Proceedings of the Natural Legal Language Processing Workshop 2021**. Punta Cana, Dominican Republic: Association for Computational Linguistics, nov. 2021. P. 210–216. DOI: 10.18653/v1/2021.nllp-1.22. Disponível em: <https://aclanthology.org/2021.nllp-1.22>. Citado na p. 38.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to Information Retrieval**. Cambridge University Press, 2009. Citado na p. 21.

MEDIUM. **Natural Language Processing (NLP) based Chatbots**. Acesso em 05 dez. 2023. 2023. Disponível em: <https://medium.com/analytics-vidhya/natural-language-processing-nlp-based-chatbots-7b2436428256>. Citado nas pp. 20, 24.

NAKAMURA, A. L. d. S. A infraestrutura e a corrupção no Brasil Infrastructure and corruption in Brazil. **Revista Brasileira de Estudos Políticos**, n. 17, p. 97–126, 2018. Citado na p. 12.

NAVIGLI, R. **Lecture 7: Word Sense Disambiguation**. Acesso em 04 dez. 2023. 2013. Disponível em: <https://naviglinlp.blogspot.com/2013/05/lecture-7-word-sense-disambiguation.html>. Citado na p. 22.

NETRIN. **Web Scraping: O que é, como funciona e para que serve?** 2023. Disponível em: <https://netrin.com.br/web-scraping-o-que-e-como-funciona/>. Acesso em: 1 set. 2023. Citado na p. 45.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up?: sentiment classification using machine learning techniques. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. 2002. P. 79–86. Citado na p. 23.

PORTAL DA TRANSPARÊNCIA. **Ação Orçamentária no Enfrentamento da Emergência de Saúde Pública de Importância Internacional Decorrente do Coronavírus**. 2023. 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: 2 ago. 2023. Citado na p. 12.

PRAKASH, A. **What is transformer architecture and how does it power ChatGPT?** 2023. https://www.thoughtspot.com/data-trends/ai/whatis-transformer-architecture-chatgpt. Acesso em: 29 ago. 2023. Citado na p. 13.

REITER, E.; DALE, R. Building natural language generation systems. **Artificial Intelligence**, Elsevier, v. 125, n. 1-2, p. 161–188, 2000. Citado na p. 23.

SALAS, J.; BARROS VIDAL, F. d.; 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.

SANH, V.; DEBUT, L.; CHAUMOND, J.; WOLF, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. **ArXiv**, abs/1910.01108, 2019. Citado nas pp. 37, 50.

SMITH, J. A.; JOHNSON, B. D. Detecting Bid Rigging in Procurement Auctions Using Machine Learning. **Journal of Public Administration Research and Theory**, v. 28, n. 3, p. 414–429, 2018. Citado na p. 12.

SPACY. **Spacy: Linguistic Features**. Acesso em 04 dez. 2023. 2023. Disponível em: <https://spacy.io/usage/linguistic-features>. Citado nas pp. 21, 23.

THIRDEYE. **OpenAI Application for Sentiment Analysis**. Acesso em 04 dez. 2023. 2023. Disponível em: <https://thirdeyedata.ai/open-ai-applications/sentiment-analysis/>. Citado na p. 23.

TJONG KIM SANG, E. F.; DE MEULDER, F. Introduction to the CoNLL-03 shared task: Language-independent named entity recognition. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. 2003. P. 142–147. Citado na p. 22.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. **Attention Is All You Need**. 2017. arXiv: 1706.03762 [cs.CL]. Citado nas pp. 25, 33, 37.

VATSAL, S.; MEYERS, A.; ORTEGA, J. E. **Classification of US Supreme Court Cases using BERT-Based Techniques**. 2023. arXiv: 2304.08649 [cs.CL]. Citado na p. 38.

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. v.; 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: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020. P. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. Disponível em: <https://aclanthology.org/2020.emnlp-demos.6>. Citado na p. 13.

ZHANG, M.; ZHOU, W. Fraud Detection in Public Procurement: A Network Analysis Approach. **Decision Support Systems**, v. 137, p. 113365, 2020. Citado na p. 12.

ZHANG, Z.; ZHANG, Z.; CHEN, H.; ZHANG, Z. A Joint Learning Framework With BERT for Spoken Language Understanding. **IEEE Access**, v. 7, p. 168849–168858, 2019. DOI: 10.1109/ACCESS.2019.2954766. Citado na p. 49.