Universidade de Brasília
Faculdade de Tecnologia

Avaliando o relacionamento semântico de textos em publicações de Diários Oficiais a partir de modelos profundos com Transformers

Francisco Henrique da Silva Costa
Isabela Maria Pereira Cruzeiro

PROJETO FINAL DE CURSO
ENGENHARIA DE CONTROLE E AUTOMAÇÃO

Brasília
2023

Universidade de Brasília
Faculdade de Tecnologia

Avaliando o relacionamento semântico de textos em publicações de Diários Oficiais a partir de modelos profundos com Transformers

Francisco Henrique da Silva Costa
Isabela Maria Pereira Cruzeiro

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

Silva Costa, Francisco Henrique da.
S586a Avaliando o relacionamento semântico de textos em publicações de Diários Oficiais a partir de modelos profundos com Transformers / Francisco Henrique da Silva Costa; Isabela Maria Pereira Cruzeiro; orientador Flávio de Barros Vidal. -- Brasília, 2023.
69 p.

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

1. Transformers. 2. Relacionamento semântico. 3. Entidades contratuais. 4. Processamento de Linguagem Natural. I. Pereira Cruzeiro, Isabela Maria. II. Vidal, Flávio de Barros, orient. III. Título

Universidade de Brasília
Faculdade de Tecnologia

Avaliando o relacionamento semântico de textos em publicações de Diários Oficiais a partir de modelos profundos com Transformers

Francisco Henrique da Silva Costa
Isabela Maria Pereira Cruzeiro

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

Trabalho aprovado. Brasília, 14 de Dezembro de 2023:

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

Prof. Dra. Carla Maria C. e C. Koike,
UnB/IE/CIC
Examinador interno

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

Brasília
2023

Eu dedico este trabalho as pessoas que contribuíram com minha vida acadêmica, em especial meus pais e minha irmã, que contribuíram para a pessoa que sou hoje.

Francisco Henrique da Silva Costa

Este trabalho é dedicado a todos que participaram da minha trajetória de alguma forma, sobretudo aos meus pais, que tanto fizeram por mim ao longo de todos esses anos.

Isabela Maria Pereira Cruzeiro

Agradecimentos

Eu agradeço novamente meus pais Francisco e Veridiana e minha irmã Isa por serem minha fonte inesgotável de encorajamento e por sempre estarem ao meu lado, me apoiando nos meus momentos mais difíceis.

Gostaria de agradecer todos os meus colegas que participaram na minha vida acadêmica, independente se continuam comigo até hoje no curso e participando dos diversos trabalhos em grupo juntos, ou se acabaram por escolher outros caminhos na vida. É para essas pessoas eu gostaria de falar que nós somos sempre mais fortes quando estamos juntos.

Gostaria de destacar um agradecimento para a Isabela por ser uma pessoa incrível e me convidar para essa jornada que juntos enfrentamos desafios e compartilhamos ideias, permitindo por fim formar esse trabalho juntos.

Por fim eu gostaria de agradecer o Prof. Dr. Flavio de Barros Vidal pela quantidade inigualável de apoio e dedicação que ele nos deu afim de produzir o trabalho no melhor ambiente possível.

Francisco Henrique da Silva Costa

Em primeiro lugar agradeço meus pais, Margareth e Paulo por me darem apoio incondicional em toda a minha trajetória na Universidade, em todos os momentos, sem eles eu não teria conquistado tudo que conquistei. Agradeço minha irmã, Mayra, que é a melhor amiga que eu poderia ter, e que tanto me deu apoio em todos os momentos difíceis, essa conquista é um pouco dela também.

Agradeço a todos os meus colegas de curso, até mesmo aqueles que ficaram pelo caminho para trilhar outros caminhos, em todos os trabalhos em grupo que fizemos em conjunto, todas as listas de exercícios e todo o conhecimento que aprendi com vocês, obrigada pelos anos de faculdade. Destaco aqui, minha dupla no trabalho de graduação, Francisco, que tanto me ajudou e colaborou para a realização desse trabalho.

Por fim, agradeço ao Prof. Dr. Flavio de Barros Vidal que ao longo de um ano de trabalho teve muita paciência para nos ensinar e orientar para realizar o trabalho da melhor maneira possível. Toda a dedicação conosco será sempre lembrada.

Isabela Maria Pereira Cruzeiro

Resumo

A comunicação oficial do país é conduzida por meio de diários oficiais, que apresenta grande volume de publicações em formato textual, neste ambiente existe o problema das fraudes em licitações públicas que gera prejuízos aos cofres públicos e danos à sociedade. A identificação de práticas corruptivas em documentos publicados nesse meio é desafiadora em termos de acesso e organização. Nesse sentido, o trabalho proposto visa desenvolver metodologias para a extração e organização de informações presentes nos diários oficiais, fornecendo suporte aos tomadores de decisão no combate à corrupção e fraudes em instrumentos contratuais publicados na Seção 3 do Diário Oficial da União. Por meio de técnicas de processamento de linguagem natural com *Transformers*, é esperado que seja possível definir o relacionamento semântico entre entidades contratuais que exigem publicação, para facilitar a visualização de documentos correlacionados para o especialistas. Um conjunto de dados foi criado com publicações da Seção 3 do DOU para avaliar a semelhança entre publicações de um mesmo convênio. Realizou-se um levantamento bibliográfico e definição de tarefas, como *Passage Ranking* e *Semantic Textual Similarity*, que fundamentaram o estudo. Por meio dos métodos adotados, foram conduzidas análises e estudos de *Sentence-Transformers* para a seleção do modelo a ser treinado. Os resultados indicam que o modelo treinado alcançou uma acurácia de 96%, representando um avanço em relação ao modelo pré-treinado. No entanto, a precisão, a sensibilidade e o *F1-score* não demonstraram melhoria no desempenho.

Palavras-chave: Transformers, Relacionamento semântico, Entidades contratuais, Processamento de Linguagem Natural.

Abstract

The official communication in Brazil is conducted through official journals, which present a high volume of publications daily. In this environment, there is the issue of fraud in public tenders that result in losses to public funds and harm to society. Identifying corrupt practices in documents published in this medium is challenging regarding access and organization. In this regard, the proposed work aims to develop methodologies for the extraction and organization of information present in official journals, supporting decision-makers in the fight against corruption and fraud in contractual instruments published in Section 3 of Diário Oficial da União (DOU). Through natural language processing techniques with Transformers, it is expected that it will be possible to define the semantic relationship between contractual entities that require publication to facilitate the visualization of correlated documents for experts. A data set was created with publications from Section 3 of the DOU to evaluate the similarity between publications from the same agreement. A bibliographic survey and definition of techniques, such as Passage Ranking and Semantic Textual Similarity, were carried out, which are the study's baseline. Analyses and studies of Sentence Transformers are used through the selected methods to select the model to be trained. The results indicate that the trained model achieved an accuracy of 96%, representing an improvement over the pre-trained model. However, Precision, Recall, and F1-score did not demonstrate performance improvement.

Keywords: Transformers, Semantic similarity , Contractual entities, Natural Language Processing.

Lista de ilustrações

Figura 2.1 – Etapas do convênio até contrato. 17
Figura 2.2 – Exemplo de publicação de convênio. 18
Figura 2.3 – Exemplos de publicação no DOU de licitações. 19
Figura 2.4 – Exemplo de publicação de contrato. 20
Figura 2.5 – Exemplo de publicação de termo aditivo. 21
Figura 2.6 – Exemplo de publicação de prorrogação de ofício. 21
Figura 2.7 – Organização *Transformer Block*. 26
Figura 2.8 – Fluxo de informação da camada *self-attention*. 27
Figura 2.9 – Fluxo de informação da camada *self-attention*. 29
Figura 2.10–Organização do *Feedfoward*. 31
Figura 2.11–Gráfico para o cálculo da EER. 34
Figura 4.12–Metodologia proposta. 38
Figura 4.13–Connected Papers. 40
Figura 4.14–Descrição da publicação. 42
Figura 4.15–Conjunto de dados para o treinamento. 44
Figura 4.16–Processo Validação Cruzada. 45
Figura 5.17–Estrutura do conjunto de dados em JSON. 48
Figura 5.18–Representação em Boxplot da quantidade de (a) caracteres , (b) palavras e (c) conjunto de convênio. 49
Figura 5.19–Funcionamento dos modelos de similaridade semântica. 50
Figura 5.20–Gráficos de EER para a (a) *Passage Ranking* e (b) *Semantic Textual Similarity*. 52
Figura 5.21–Publicações de convênios distintas. 60

Lista de tabelas

Tabela 2.1 – Matriz de Confusão. 32
Tabela 3.2 – Comparação modelos *sentence-transformers*. 36
Tabela 5.3 – Descrição do conjunto de dados. 48
Tabela 5.4 – Estatísticas do conjunto de dados. 48
Tabela 5.5 – Estatísticas do conjunto de convênios do conjunto de dados. 49
Tabela 5.6 – Acurácia Top-N para modelos de referência 51
Tabela 5.7 – Resultados com a base de 559 publicações. 51
Tabela 5.8 – Resultados modelos em inglês. 53
Tabela 5.9 – Resultados modelos em português. 53
Tabela 5.10–Métricas de avaliação modelos pré-treinados. 54
Tabela 5.11–Métricas de avaliação nos modelos pré-treinados em língua portuguesa. 54
Tabela 5.12–Resultados do treinamento com 3 iterações. 56
Tabela 5.13–Resultados do treinamento com 5 iterações. 56
Tabela 5.14–Resultados do treinamento com 10 iterações. 56

Lista de abreviaturas e siglas

BERT	Bidirectional Encoder Representations for Transformers	35
DNN	Deep Neural Network	35
DOU	Diário Oficial da União	16
EER	Equal Error Rate	34
ELMO	Embeddings from Language Models	23
FAR	False Acceptance Rate	34
FN	False Negative	32
FNN	Feed Foward Network	29
FP	False Positive	32
FRR	False Rejection Rate	34
GLOVE	Global Vectors for Word Representation	23
GPT	Generative Pre-trained Transformer	42
LSTM	Long Short-Term Memory	35
MT	Machine Translation	22
NER	Named Entity Recognition	23
NLP	Natural Language Processing	22
RoBERTa	Robustly Optimized BERT Pretraining Approach	35
SBERT	Sentence-BERT	35
STS	Semantic Textual Similarity	23
TN	True Negative	32
TP	True Positive	32

Sumário

1 Introdução 13
1.1 Motivações e justificativas 13
1.2 Semântica em Elementos Textuais 14
1.3 Objetivos geral e específicos 15
1.3.1 Objetivo Geral 15
1.3.2 Objetivos Específicos 15
1.4 Organização do Trabalho 15

2 Fundamentação teórica 16
2.1 Diário Oficial da União 16
2.2 Convênios, Licitação, Contratos 16
2.2.1 Convênios 17
2.2.2 Licitações 18
2.2.3 Contratos 19
2.2.4 Termo Aditivo 20
2.2.5 Prorrogação de Ofício 21
2.3 Processamento de linguagem natural 22
2.4 Similaridade Semântica Textual 23
2.4.1 Embeddings 23
2.4.2 Métricas de Similaridade 24
2.5 Transformers 25
2.5.1 Transformer Block 25
2.5.2 Self-Attention 26
2.5.3 Conexão residual e Normalização 29
2.5.4 Feedfoward 31
2.6 Métricas de avaliação 32
2.6.1 Acurácia Top-N 33
2.6.2 Equal Error Rate 33

3 Trabalhos relacionados 35
3.1 Uso de Modelos com *Transformers* 35
3.2 Similaridade Semântica 36

4 Metodologia proposta 38
4.1 Levantamento Bibliográfico 38
4.2 Conjunto de Dados 41

4.3 Métodos de Referência . . . 42
4.4 Avaliação e Escolha de Modelos . . . 43
4.5 Treinamento . . . 43
4.5.1 Conjunto de dados do treinamento . . . 43
4.5.2 Validação Cruzada . . . 44
4.5.3 Função de perda . . . 45
4.5.4 Avaliação do Treinamento . . . 46
4.6 Avaliação dos Resultados . . . 46

5 Resultados . . . 47
5.1 Levantamento Bibliográfico . . . 47
5.2 Conjunto de Dados . . . 47
5.3 Métodos de Referência . . . 50
5.4 Análise e Estudo de Modelos . . . 52
5.5 Treinamento . . . 54
5.6 Análise de resultados . . . 56

6 Conclusões . . . 61

Referências . . . 63

1 Introdução

1.1 Motivações e justificativas

A corrupção é apontada como uma grande ameaça a democracia com consequências não apenas políticas como também sociais. No Brasil, diversos escândalos políticos são causados por desvios de recursos públicos, uso indevido da máquina administrativa, sistemas de propinas entre diversas outras formas de corrupção, fraudes e licitações ilícitas (TEIXEIRA; REHBEN-SATLHER; RODRIGUES, 2021).

Rente Neto (2021) analisou as práticas corruptivas mais frequentes no período de 2004 a 2015 através de denúncias apuradas com a aceitação da denúncia ou pelo arquivamento no Supremo Tribunal Federal, onde determinou que 57,6% das denúncias foram de crimes de lavagem de dinheiro público e formação de quadrilha, logo em seguida, vem a prática de irregularidades na licitação, representando 13,7% das denúncias e por fim, 6,8% das denúncias vem da captação ou uso ilícito de recursos eleitorais com 6,8%.

As licitações irregulares trazem prejuízos aos cofres públicos e danos à sociedade, onde suspeita-se de contratações de obras e serviços, compras de equipamentos e materiais de expediente e sobre processos de escolhas de construtores, prestadores de serviços e fornecedores. A nova Lei de Licitações e Contratos (14.133/2021) visa a modernização e tornar mais transparentes e seguros na esfera jurídica todo esse processo licitatório (OLIVEIRA, 2021).

O combate à corrupção requer uma série de medidas cautelosas. No cenário de licitações e contratações públicas, é essencial a existência de sistemas que não apenas desestimulem e reduzam práticas ilícitas, mas também detectem e punam tais ações. Esses sistemas contribuem no compromisso da sociedade e do governo em abordar esta questão, como apresentado em Fortini e Motta (2003).

De acordo com STF (2023), todo cidadão tem direito a informação das decisões do Estado e deve ser assegurado o acesso às decisões públicas e atividades dos servidores públicos. O art. 5º, inciso XXXIII da Constituição Federal refere-se:

> "Todos têm direito a receber dos órgãos públicos informações de seu interesse particular, ou de interesse coletivo ou geral, que serão prestadas no prazo da lei, sob pena de responsabilidade, ressalvadas aquelas cujo sigilo seja imprescindível à segurança da sociedade e do Estado." (BRASIL, 1988)

O princípio da publicidade é uma forma de garantir a transparência, a moralidade e o controle da atividade administrativa pelo povo. Conforme explicitado em STF (2023),

todos os atos e decisões da Administração Pública devem ser publicados de forma clara e completa, salvo aqueles que envolvam sigilo imprescindível à segurança da sociedade e do Estado.

Dessa forma, o Decreto 9.215 (BRASIL, 2017a), regulamenta a publicação de toda a comunicação dos atos do governo federal no Diário Oficial da União, que deve ser realizada diariamente de segunda-feira a sexta-feira. Devem ser publicados todos os atos com conteúdo normativo, atos oficiais de toda a administração pública e também todos os atos que envolvem procedimentos de licitações, extratos de instrumentos contratuais e editais.

Baseado nestas prerrogativas, em Lima et al. (2020) é proposta uma metodologia para a realização de detecção e classificação de padrões de conluio em licitações públicas. Nesse contexto, o trabalho proposto visa auxiliar no desenvolvimento de metodologias de técnicas de combate e detecção de fraudes advindas dos extratos de instrumentos contratuais publicados na Seção 3 do Diário Oficial da União (DOU). Por meio de técnicas de processamento de linguagem natural com transformadores (auto-supervisionados ou não supervisionados), espera-se ser possível aprimorar o desenvolvimento atual para definição do relacionamento semântico entre entidades contratuais que exigem publicação, de acordo com obrigações legais ou regulamentares, facilitando a visualização de documentos correlacionados para os especialistas.

1.2 Semântica em Elementos Textuais

Em linguística, **semântica** é o estudo do significado das palavras, frases, sentenças dentro de um determinado contexto. A semântica é uma das principais ciências da linguística, em conjunto com a fonética, fonologia, morfologia e sintaxe (CANÇADO, 2008).

A semântica estuda o entendimento da utilização de palavras e expressões para transmitir a informação e representar ideias e conceitos, analisando a relação entre as palavras e os diferentes significados que uma palavra pode assumir em diferentes contextos e como esses significados relacionam-se formando um conjunto entendível de texto (CANÇADO, 2008).

Alguns aspectos importantes abordados pela semântica são: a denotação das palavras, ou seja, o significado, literalidade das palavras e expressões, conotação, o sentido figurado daquela expressão, a sinonímia e antonímia, que é o estudo da equivalência e da oposição das palavras, a polissemia, onde mostra que palavras e expressões possuem multiplicidade de sentidos, homonimia que refere-se a palavras iguais com significado dissemelhante e ambiguidade, ou seja a duplicidade de interpretação de uma expressão em determinado contexto (MORAES PINTO et al., 2016).

Dessa forma, a semântica é de extrema relevância para a compreensão da linguagem,

porque determina a forma que palavras, sentenças e estruturas linguísticas são utilizadas para expressar significados e ideias (CANÇADO, 2008).

A semelhança semântica textual diz respeito à similaridade de significado entre diferentes partes de um texto. Conforme destacado por Ramalho (2008), a estruturação sintática de uma frase não é suficiente para a determinação do sentido de um fragmento de texto. É importante considerar os elementos semânticos para a transmissão adequada de significados. Desta maneira, estruturas sintáticas não necessariamente denotam o mesmo sentido, mas sim a forma de transmissão de significados dentro do contexto linguístico.

1.3 Objetivos geral e específicos

1.3.1 Objetivo Geral

O objetivo do presente trabalho é realizar o treinamento de modelos de *Transformers* previamente treinados para a Semelhança Semântica Textual. Sendo sua principal atividade o estabelecimento de relação semântica entre as elementos de publicações contratuais do Diário Oficial da União, comparando os diferentes modelos de *Transformers* disponíveis na literatura e avaliar o desempenho no processo de detecção e classificação semântica textual.

1.3.2 Objetivos Específicos

Os objetivos específicos a serem alcançados neste trabalho, são:

- Revisão da literatura sobre o tema de relacionamento semântico em textos com técnicas de processamento de linguagem natural.
- Elaboração do conjunto de dados de publicações anotadas, estabelecendo uma hierarquia de relacionamentos semântico entre elas.
- Análise e avaliação dos modelos profundos que utilizam *Transformer* para avaliação da relação semântica textual entre as publicações anotadas.

1.4 Organização do Trabalho

O Capítulo 2 é feita a fundamentação teórica utilizada e as definições de *Transformers*, Similidade Semântica. O Capítulo 3 foi realizada a coleta dos trabalhos correlacionados no contexto acadêmico. O Capítulo 4 será descrita toda a metodologia empregada. O Capítulo 5 é feita a discussão de análises e resultados obtidos. Por fim, o Capítulo 6 é concluído o trabalho apresentando as considerações finais.

2 Fundamentação teórica

O presente Capítulo estabelece os fundamentos teóricos que servem de base para o desenvolvimento e as análises subsequentes. Neste sentido, será abordada a estrutura das publicações a serem avaliadas, como também a organização do Diário Oficial da União. Serão discutidas as principais técnicas de Processamento de Linguagem Natural e os métodos para determinação de similaridade semântica. Por fim, serão apresentadas as métricas de avaliação dos modelos utilizados.

2.1 Diário Oficial da União

O Diário Oficial da União possui 160 anos de existência e teve sua primeira publicação sob o nome de "Diario Official do Império do Brasil", quando, em outubro de 1862, foi impresso e distribuído pela primeira vez. A relevância desse jornal é ressaltada pelo Artigo 37 da Constituição da República de 1988, que enfatiza o princípio da publicidade dos atos administrativos públicos. Desde 2017, o DOU deixou de ser veiculado em formato impresso, estando disponível exclusivamente em meio eletrônico (BRASIL, 2017b).

O DOU é o veículo de comunicação oficial do Governo Federal para tornar público todo assunto de âmbito federal. No jornal é publicado decisões, ações, resoluções do governo e a divulgação de informações de interesse público, como leis, decretos, contratos, editais e etc (MACÊDO, 2018).

De acordo com Brasil (2023a), as publicações no DOU são divididas em três seções:

- **Seção 1**: publicações de normas, atos normativos, leis, resoluções, decretos, portarias.
- **Seção 2**: publicações relativas à pessoal, como nomeações, designações de cargos comissionados.
- **Seção 3**: publicações de contratos, convênios, editais, comunicados e entre outros.

2.2 Convênios, Licitação, Contratos

A estrutura de dados classificados no presente trabalho é organizada de acordo com os autos presentes em um processo licitatório definido pela Lei 14.133/2021. O processo licitatório visa assegurar a contratação mais vantajosa para a Administração Pública, promover a justa competição e evitar contratações com superfaturamento na execução de contratos (BRASIL, 2021).

A Figura 2.1 ilustra a forma como as publicações são relacionadas/estruturadas por meio de etapas dentro de um processo licitatório. Após o ajuste celebrado entre a administração pública e a convenente, ocorre a publicação no Diário Oficial da União (DOU), seguida pela publicação do edital de licitação, que pode resultar em um ou mais contratos para a realização de obras, bens e serviços. Esses contratos e convênios podem sofrer alterações de vigência, cláusulas e aspectos financeiros, sendo denominados termos aditivos.

Figura 2.1 – Etapas do convênio até contrato.

Fonte: (DEEP VACUITY, 2021)

2.2.1 Convênios

Os convênios são celebrados entre entidades de Administração Pública ou organizações não -governamentais para a transferências de recursos financeiros para a consecução de um objetivo comum. O objeto, o produto do convênio, pode envolver realização de projetos, atividades, serviço e aquisição de bens (PORTAL DA TRANSPARÊNCIA, 2023a).

As partes envolvidas, ou partícipes são os instrumentos jurídicos utilizados, sendo eles;

- Concedente: órgão que repassa o recurso.
- Convenente: entidade a qual Administração Pública repassa o recurso.

DIÁRIO OFICIAL DA UNIÃO

Publicado em: 22/01/2019 | Edição: 15 | Seção: 3 | Página: 95

Órgão: Ministério da Saúde/Gabinete do Ministro

EXTRATO DE CONVÊNIO

Espécie: Convênio Nº 882484/2019, Nº Processo: 25000226133201872, Concedente: MINISTERIO DA SAUDE, Convenente: MISSAO EVANGELICA CAIUA CNPJ nº 03747268000180, Objeto: PRESTAÇÃO DE SERVIÇOS E AÇÕES COMPLEMENTARES NA ÁREA DE ATENÇÃO À SAÚDE, VISANDO O ATINGIMENTO DOS OBJETIVOS ESPECÍFICOS ESTABELECIDOS PELA SECRETARIA ESPECIAL DE SAÚDE INDÍGENA - SESAI EM CONSONÂNCIA COM A POLÍTICA NACIONAL DE ATENÇÃO À SAÚDE DOS POVOS INDÍGENAS - PNASPI E AS ESPECIFICIDADES SÓCIO-CULTURAIS DOS POVOS INDÍGENAS, NO ÂMBITO DO SUBSISTEMA DE ATENÇÃO À SAÚDE INDÍGENA - SASISUS (CHAMADA PÚBLICA 11/2018), Valor Total: R$ 20.150.661,00, Valor de Contrapartida: R$ 0,00, Valor a ser transferido ou descentralizado por exercício: 2019 - R$ 14.105.462,00; 2020 - R$ 6.045.199,00, Crédito Orçamentário: Num Empenho: 2019NE800021, Valor: R$ 4.030.132,00, PTRES: 109694, Fonte Recurso: 6151000000, ND: 33504305, Vigência: 17/01/2019 a 17/01/2020, Data de Assinatura: 17/01/2019, Signatários: Concedente: LUIZ HENRIQUE MANDETTA CPF nº 519.421.431-68, Convenente: SILAS DE SOUZA DA SILVA CPF nº 421.708.001-82.

Figura 2.2 – Exemplo de publicação de convênio.

Fonte: (IMPRENSA NACIONAL, 2023)

2.2.2 Licitações

A licitação é o processo por meio do qual a Administração Pública contrata obras, serviços, compras, alienação e locações (PORTAL DA TRANSPARÊNCIA, 2023b).

Segundo o Art. 11 da Lei 14.133/2021 (BRASIL, 2021), o processo licitatório tem como objetivo assegurar a seleção da proposta apta a gerar o resultado de contratação mais vantajoso para a Administração Pública, assegurar tratamento igual entre os licitantes, evitar contratações com sobrepreço e incentivar a inovação e o desenvolvimento nacional sustentável.

O Art. 28 da Lei 14.133/2021 define as seguintes modalidades de licitação:

a) Pregão: modalidade destinada à contratação de bens e serviços comuns.

b) Concorrência: modalidade destinada para contratação de bens e serviços especiais e de obras de serviço comum.

c) Concurso: modalidade para escolha de trabalho técnico, cientifico ou artístico.

d) Leilão: modalidade para aquisição de bens a quem oferecer o maior lance.

e) Diálogo competitivo: modalidade em que há o diálogo com o licitante selecionado previamente mediante critérios.

DIÁRIO OFICIAL DA UNIÃO

Publicado em: 29/04/2022 | Edição: 80 | Seção: 3 | Página: 257

Órgão: Governo do Estado/Governo do Estado de Santa Catarina/FUNDO PENITENCIÁRIO DO ESTADO DE SANTA CATARINA/SECRETARIA DE ESTADO DA ADMINISTRAÇÃO PRISIONAL E SOCIOEDUCATIVA

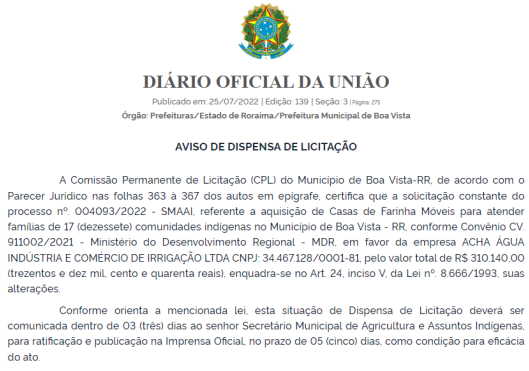
AVISO DE LICITAÇÃO

PREGÃO ELETRÔNICO Nº 71/SAP/2022

A Secretaria de Estado da Administração Prisional e Socioeducativa comunica Pregão Eletrônico nº 071/2022 - menor preço por lote.

Pregão Eletrônico nº 0071/2022 - menor preço por Lote. Objeto: Aquisição de equipamentos e insumos às Oficinas Laborais de Artefatos de Cimento da Penitenciária Agrícola de Chapecó e Penitenciária da Região de Curitibanos - Projeto PROCAP 5º Ciclo - Convênio SICONV nº 891728/2019. Início da entrega de propostas: às 13:10 horas do dia 02/05/2022. Fim da entrega de propostas: às 13:15 horas do dia 12/05/2022. Abertura da sessão: a partir das 13:15 horas do dia 12/05/2022. Início da disputa: a partir das 13:30 horas do dia 12/05/2022. O Edital e seus anexos estão disponíveis no site www.sap.sc.gov.br. Informações sobre o edital serão prestadas através do e-mail geliclicitacao@sap.sc.gov.br, no horário das 12:00 às 19:00, em dias úteis. Processo SGP-e: SAP 00111165/2021. GGG: 2022AS004388.E-Sfinge: 1D32EBE121F5E671DC94BA909A76F191A23713F7

(a) Aviso de Licitação.



DIÁRIO OFICIAL DA UNIÃO

Publicado em: 25/07/2022 | Edição: 139 | Seção: 3 | Página: 275

Órgão: Prefeituras/Estado de Roraima/Prefeitura Municipal de Boa Vista

AVISO DE DISPENSA DE LICITAÇÃO

A Comissão Permanente de Licitação (CPL) do Município de Boa Vista-RR, de acordo com o Parecer Jurídico nas folhas 363 à 367 dos autos em epígrafe, certifica que a solicitação constante do processo nº. 004093/2022 - SMAAI, referente a aquisição de Casas de Farinha Móveis para atender famílias de 17 (dezessete) comunidades indígenas no Município de Boa Vista - RR, conforme Convênio CV. 911002/2021 - Ministério do Desenvolvimento Regional - MDR, em favor da empresa ACHA ÁGUA INDÚSTRIA E COMÉRCIO DE IRRIGAÇÃO LTDA CNPJ: 34.467.128/0001-81, pelo valor total de R$ 310.140,00 (trezentos e dez mil, cento e quarenta reais), enquadra-se no Art. 24, inciso V, da Lei nº. 8.666/1993, suas alterações.

Conforme orienta a mencionada lei, esta situação de Dispensa de Licitação deverá ser comunicada dentro de 03 (três) dias ao senhor Secretário Municipal de Agricultura e Assuntos Indígenas, para ratificação e publicação na Imprensa Oficial, no prazo de 05 (cinco) dias, como condição para eficácia do ato.

(b) Aviso de Dispensa de Licitação.

Figura 2.3 – Exemplos de publicação no DOU de licitações.

Fonte: (IMPRENSA NACIONAL, 2023)

2.2.3 Contratos

Os contratos são todo e qualquer ajuste firmado entre a Administração Pública e a empresa vencedora do processo licitatório que estabelecem com clareza e precisão as condições de execução por meio de clausulas definindo os direitos, obrigações e responsabilidades das partes a parte do processo licitatório vinculado (TRIBUNAL DE CONTAS DA UNIÃO, 2010).

De acordo com a Lei 14.133/2021, todo contrato deve mencionar os nomes das partes e de representantes, a finalidade, o ato de autorização, o número do processo de licitação e as cláusulas contratuais. Em todo contrato é estabelecido o objeto de contrato, a vinculação ao edital de licitação, a legislação aplicável, o regime de execução ou fornecimento, o preço e a condições de pagamento, os prazos de início, execução, conclusão, entrega e pagamento, a obrigação do contratado de manter as obrigações assumidas, e etc (BRASIL, 2021).

DIÁRIO OFICIAL DA UNIÃO

Publicado em: 05/05/2020 | Edição: 84 | Seção: 3 | Página: 190

Órgão: Prefeituras/Estado do Paraná/Prefeitura Municipal de Pérola

EXTRATO DE CONTRATO

Contrato de Fornecimento nº 35/2020.

Pregão Presencial nº 08/2020

Contratante: MUNICÍPIO DE PÉROLA

Contratada: SCHLICKMANN & ROTTA LTDA-ME

Objeto: Aquisição de equipamentos agrícola novos, com recursos provenientes do convênio MAPA nº 889856/2019, firmado entre o Ministério da Agricultura, Pecuária e Abastecimento-MAPA, e o Município de Pérola, Estado do Paraná.

Valor Total: R$ 22.280,00 (vinte dois mil duzentos e oitenta reais)

Vigência: 29/04/2020 a 29/04/2021.

Adjudicada e Homologada: 24/04/2020.

Figura 2.4 – Exemplo de publicação de contrato.

Fonte: (IMPRENSA NACIONAL, 2023)

2.2.4 Termo Aditivo

O aditivo contratual é um complemento ao contrato assinado anteriormente. É realizado na alteração de qualquer umas das cláusulas como por exemplo alteração de prazos, financeiros ou objeto. É comum os convênios e contratos possuírem termos aditivos vinculados (DIRETORIA DE COMPRAS, CONTRATOS E CONVÊNIOS , 2018).

Nos convênios, o termo aditivo é a alteração de qualquer das cláusulas do convênio de adesão celebrado entre a concedente e convenente (BRASIL, 2023b).

DIÁRIO OFICIAL DA UNIÃO

Publicado em: 06/06/2022 | Edição: 106 | Seção: 3 | Página: 144

Órgão: Ministério da Justiça e Segurança Pública/Secretaria Nacional de Justiça

EXTRATO DE TERMO ADITIVO

Espécie: Termo Aditivo de Alteração da Vigência/ Acréscimo Nº 000001/2022 ao Convênio Nº 902184/2020. Convenentes: Concedente: MINISTERIO DA JUSTICA E SEGURANCA PUBLICA, Unidade Gestora: 200143. Convenente: MINISTERIO PUBLICO DO ESTADO DO AMAPA, CNPJ nº 34869354000199. Prorrogação do prazo de vigência até 03 de dezembro de 2022, acréscimo de contrapartida e ajuste ao Plano de Trabalho ao Termo de Convênio Plataforma +BRASIL nº 902184/2020. Valor Total: R$ 178.620,67, Valor de Contrapartida: R$ 178.620,67, Vigência: 04/06/2022 a 03/12/2022. Data de Assinatura: 03/12/2020. Signatários: Concedente: BRUNO ANDRADE COSTA, CPF nº 88643727172, Convenente: IVANA LUCIA FRANCO CEI, CPF nº 223.200.242-04.

Figura 2.5 – Exemplo de publicação de termo aditivo.

Fonte: (IMPRENSA NACIONAL, 2023)

2.2.5 Prorrogação de Ofício

A Prorrogação de Ofício é a prorrogação da vigência do convênio quando houver atraso na liberação de recursos por parte da Concedente (DEPARTAMENTO DE LOGÍSTICA E SERVIÇOS GERAIS, 2013).

DIÁRIO OFICIAL DA UNIÃO

Publicado em: 09/05/2023 | Edição: 87 | Seção: 3 | Página: 2

Órgão: Ministério da Agricultura e Pecuária/Secretaria de Inovação, Desenvolvimento Sustentável, Irrigação e Cooperativismo

EXTRATO DE PRORROGAÇÃO DE OFÍCIO

Espécie: Prorroga de Ofício Nº 00001/2023, ao Convênio Nº 908061/2020. Convenentes: Concedente: Ministério da Agricultura e Pecuária, , Unidade Gestora: 420013, Convenente: MUNICIPIO DE CARMESIA, CNPJ nº 18303172000108. P.I. 127/2008, art. 30, VI.. Valor Total: 440.000,00, Valor de Contrapartida: 201.250,00, Vigência: 31/12/2020 a 02/10/2023. Data de Assinatura: 05/05/2023. Assina: Pelo Ministério da Agricultura e Pecuária / RENATA BUENO MIRANDA - Secretária

Figura 2.6 – Exemplo de publicação de prorrogação de ofício.

Fonte: (IMPRENSA NACIONAL, 2023)

2.3 Processamento de linguagem natural

Processamento de linguagem natural (do inglês *Natural Language Processing* - NLP) é um ramo da inteligência artificial que permite computadores processar e manipular a linguagem humana. O NLP permite a interação entre computadores e linguagem natural de forma estruturada podendo assim realizar várias tarefas presentes no dia a dia, como automatização de tarefas, utilização de assistentes virtuais, otimização de busca de pesquisa, detecção de spams, traduções (SAS, 2023).

O estudo em NLP é amplo com diversas áreas de estudo, e diferentes problemas e que vêm ganhando espaço nos últimos 50 anos. O primeiro trabalho foi focado em *Machine Translation* (MT) por volta dos anos 40. Já no inicio da década de 1950 até os anos 1960 as pesquisas em NLP começaram a ganhar destaque com estudos focados em tradução de forma automática, morfologia (estudo da estrutura das palavras), sintaxe (estudo da estrutura gramatical das frases), semântica (estudo do significado das palavras e frase), além de tarefas como interpretação (JONES, 2001).

Com o avanço da tecnologia, desde os anos 1980, o campo de NLP ganhou mais impulso surgindo novas abordagens com o foco gramatical, Modelos de linguagem e algoritmos supervisionados e não supervisionados conforme apresentado em Jones (2001). Nos últimos anos com o aumento da capacidade computacional e advento das redes neurais, avanços significativos foram obtidos nos estudos de NLP, com novas técnicas mais avançadas sendo empregadas.

Seguindo o apresentado em (IBM, 2023), algumas das principais tarefas e aplicações das técnicas de processamento de linguagem natural são:

- Classificação de texto (*Text Classification*): processo de reconhecimento de significado de um texto de maneira o classificar. Um exemplo de aplicação é a analise de sentimentos (*Sentiment Analysis*), utilizada para classificação de textos em categorias separadas por tipo de emoção.
- Previsão de relação (*Relation Prediction*): reconhecimento de relação entre duas entidades semanticamente nomeadas.
- Reconhecimento de entidades nomeadas (*Named Entity Recognition - NER*): usada para localizar e classificar entidades em categorias pré definidas, como nome de pessoas, organizações, etc.
- Tradução (*Machine translation*): utilizada para traduzir textos em determinado idioma para outro.

- Extração de relacionamento (*Relationship Extraction*): tarefa utilizada para extrair relações semânticas em um texto. Esse relacionamento ocorre entre entidades que estão correlacionadas a categorias semânticas.
- Similaridade semântica (*Semantic Textual Similarity - STS*): determina a similaridade entre duas partes de textos.
- Normalização léxica (*Lexical Normalization*): usada para padronizar textos fora de padrão natural de linguagem.
- Respostas a perguntas (*Question answering*): tarefa utilizada para responder peguntas.
- Modelo de linguagem (*Language modeling*): utilizada para prever o próximo caractere ou próxima palavra de um texto.

2.4 Similaridade Semântica Textual

Similidade semântica entre conceitos é um método que visa medir a proximidade ou disparidade semântica entre dois conceitos de acordo com a natureza das palavras (SLIMANI, 2013).

No contexto de NLP, a comparação semântica é uma tarefa imprescindível para muitas aplicações de processamento de linguagem natural como por exemplo os mecanismos de buscas, chat bots, análise de sentimentos e outras grandes áreas de estudo e pesquisa (MAJUMDER et al., 2016).

A similaridade semântica textual (*Semantic Textual Similarity - STS*) é a tarefa de NLP que determina a semelhança entre dois textos, em sentido. Essa semelhança pode ser realizada por meio de métodos que recebem o texto como entrada e então medem a distância por meio de métricas ou por meio da conversão dos textos em vetores numéricos chamados *embeddings* que capturam a informação semântica da frase e calculam matematicamente a distância entre os vetores (WANG, J.; DONG, 2020).

2.4.1 Embeddings

Um *embedding* é uma representação numérica do texto natural para o computador entender o contexto e sentido do texto. Alguns algoritmos fazem essa transformação, como o *word2vec* (MIKOLOV et al., 2013), o GloVe (*Global Vectors for Word Representation*) (PENNINGTON; SOCHER; MANNING, 2014), o ELMo (*Embeddings from Language Models*) (PETERS et al., 2018) e o *Transformers* (VASWANI et al., 2017).

No trabalho de Mikolov et al. (2013), os autores propõe duas novas arquiteturas de modelos utilizados no processamento de linguagem natural e aprendizado de máquina, para

representação de palavras em espaço vetorial, para captar as relações semânticas entre as palavras através de representações vetoriais, onde as palavras similares possuem vetores próximos no espaço vetorial.

Já o algoritmo proposto por Pennington, Socher e Manning (2014) é realizada uma nova técnica no campo de aprendizado de máquina para representação vetorial de palavras a partir das estatísticas de co-ocorrência das palavras em um corpus, de maneira a ter informações sobre a distribuição global das palavras no texto.

Os autores em Peters et al. (2018) propõe uma representação avançada das palavras que levam em conta as características do uso das palavras como o contexto linguístico em que estão inseridas.

Por fim, conforme apresentado em Vaswani et al. (2017), os *Transformers* são arquiteturas de modelos de linguagem baseadas em *self-attention* (autoatenção), permitindo uma compreensão profunda e contextual de textos. Essa arquitetura será apresentada em detalhes na Seção 2.5.

A similaridade semântica calculada entre os *embeddings* é realizada pela medição da distância entre os vetores. Ou seja, os modelos utilizados convertem os textos nesses vetores e então calculam a proximidade da informação. As métricas de similidade mais utilizadas serão revistas a seguir na Subseção 2.4.2.

2.4.2 Métricas de Similaridade

A distância euclidiana $d(A,B)$ é uma métrica de similaridade simples que consiste na distância textual entre dois pontos no espaço Euclidiano definidos conforme a Equação 2.1 (DEZA, M. M.; DEZA, E., 2009).

$$d(A,B) = \sqrt{\sum_{i=1}^{n}(A^{(i)} - B^{(i)})^2}, \tag{2.1}$$

onde $A^{(i)}$ e $B^{(i)}$ são dois atributos e n é a dimensão da amostra.

A similidade calculada por cosseno é amplamente utilizada e é medida por meio do produto vetorial de dois vetores, onde é realizada a medição do ângulo entre dois *embeddings* e então o cosseno do ângulo é usado como a medida de similaridade entre eles como demostrado pela Equação 2.2. Se os dois dados são iguais, a similaridade será de 1, já que o ângulo é igual a 0, por exemplo, sendo assim uma métrica mais intuitiva e ao mesmo tempo custosa computacionalmente.

$$sim(A,B) = \frac{A.B}{\|A\| \, \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}, \tag{2.2}$$

onde A e B são os vetores e n é a dimensão da amostra.

2.5 Transformers

O **Transformers** é um método cada vez mais utilizado em pesquisas e projetos que envolvem Processamento de Linguagem Natural. Conforme mencionado por Jurafsky e Martin (2023), uma das principais razões para o alto desempenho desse método no entendimento da linguagem natural reside na camada de *self-attention*, que proporciona uma capacidade de abstração significativamente maior em comparação a outros modelos de aprendizado profundo. Além disso, Jurafsky e Martin (2023) destaca outra característica fundamental dos *Transformers*: a possibilidade de utilizar um modelo treinado para compreender as características linguísticas por meio do pré-treinamento e do *Transfer Learning*, o que reduz significativamente o tempo computacional do processo, tornando-o mais eficiente quando comparado a outros métodos de Processamento de Linguagem Natural (NLP). Nesta seção, será realizada uma breve explicação da fórmula de funcionamento dos *Transformers*.

2.5.1 Transformer Block

Segundo Jurafsky e Martin (2023), a arquitetura dos *Transformers* é composta por componentes denominados *Transformer block*. Esses componentes consistem em um conjunto de camadas de *self-attention*, que permite ao modelo *Transformer* abstrair o sentido do elemento textual, além de outras camadas como *feedforward*, *normalize*, e *residual connection*. A Figura 2.7 ilustra a organização desses componentes.

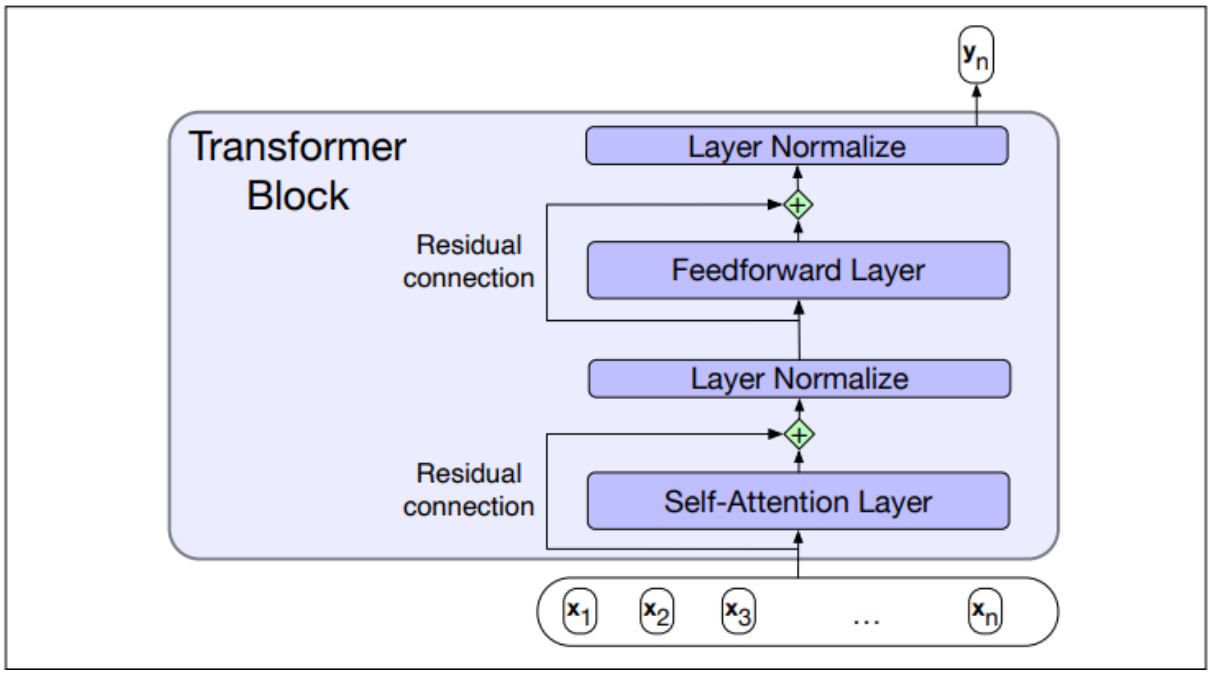


Figura 2.7 – Organização *Transformer Block*.

Fonte: (JURAFSKY; MARTIN, 2023)

Cada camada na arquitetura do modelo de *Transformers* desempenha uma função específica. Essas funções incluem otimização do treinamento para a redução de custos computacionais, prevenção da perda de informação durante a propagação entre camadas e facilitação da comunicação entre camadas e *Transformer blocks*.

2.5.2 Self-Attention

Os *Transformers* mapeiam sequências de vetores de entrada $X = [x_1, \dots, x_n]$ em sequências de vetores de saída $y_1, \dots, y_n$ do mesmo tamanho. Essa arquitetura é composta por pilhas de blocos de transformação, sendo cada bloco uma rede multi-camada que incorpora camadas lineares simples, redes *feedforward* e camadas de *self-attention*. A camada de *self-attention*, uma inovação central nos *Transformers*, possibilita a extração e utilização de informações em extensas redes de texto sem depender de conexões recursivas, como é comum em outros modelos, por exemplo, RNNs (JURAFSKY; MARTIN, 2023).

A camada de *self-attention*, como representado na Figura 2.8, realiza o mapeamento de sequências de entrada para sequências de saída de tamanho equivalente. Ao processar cada elemento do vetor de entrada, o modelo tem acesso a todas as entradas anteriores na posição correspondente do vetor, até o valor em processamento. No entanto, não possui informações sobre os itens subsequentes. É crucial notar que o cálculo é realizado independentemente para cada item, sem depender do processamento de outros itens já realizados.

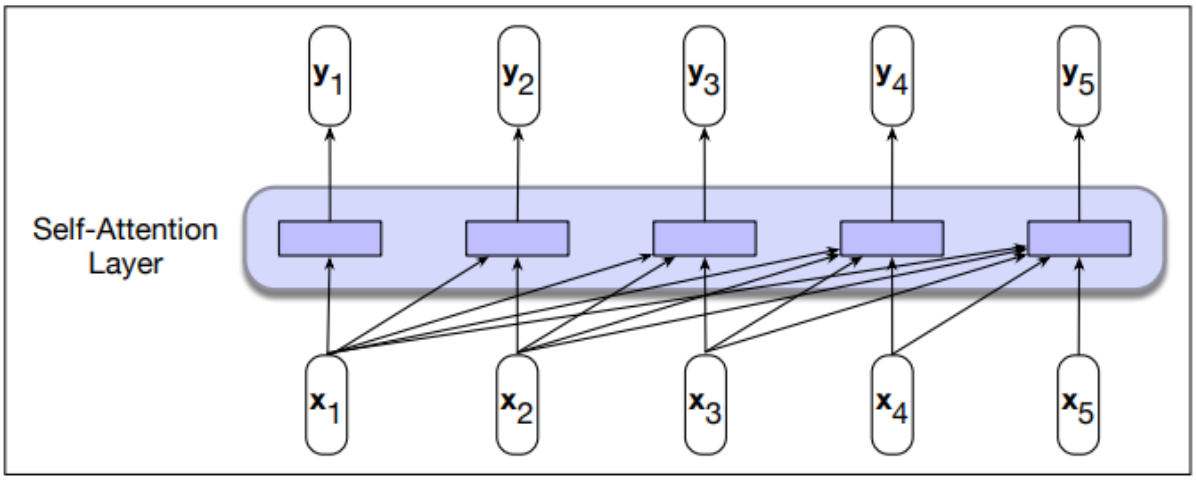


Figura 2.8 – Fluxo de informação da camada *self-attention*.

Fonte: (JURAFSKY; MARTIN, 2023)

As comparações entre os elementos realizadas na camada de *self-attention* não seguem uma sequência definida, e o resultado dessa comparação é utilizado para calcular uma saída para a entrada atual. A comparação entre elementos na camada pode ser explicada de forma simplificada por meio do produto escalar de dois valores do vetor X de entradas, gerando um escore conforme apresentado na Equação 2.3.

$$score(x_i,x_j) = x_i.x_j, \tag{2.3}$$

onde x_i e x_j são valores do vetor de entrada X nas posições i e j.

O resultado é um número, que tenta representar a relação entre os elementos do vetor X. Segundo Jurafsky e Martin (2023) para utilizar os escores, é realizada a normalização por meio da função exponencial normalizada, conhecida como *softmax*.

$$softmax(Z_i) = \frac{e^{(Z_i)}}{\sum_{j=1}^{k} e^{(Z_j)}}, \tag{2.4}$$

onde Z é o vetor a ser normalizado, sendo este vetor de tamanho K.

Com o valor normalizado pela Equação 2.4, são então calculados os valores de probabilidade α para cada combinação válida de entradas do vetor X, conforme apresentado na Equação 2.5.

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

onde x_i e x_j são valores do vetor de entrada X nas posições i e j.

A partir dos valores de probabilidades α obtidos pela Equação 2.5, é realizado o cálculo do valor de saída como explícito, na Equação 2.6.

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

onde y_i é a saída para a correspondente entrada x_i e distribuição de probabilidades α_i de tamanho j.

Os *Transformers*, de forma menos simplificada, empregam o mesmo modelo da explicação anterior com a introdução de matrizes com pesos $\mathbf{W^Q}$, $\mathbf{W^K}$ e $\mathbf{W^V}$. Esses pesos são utilizados para projetar cada vetor de entrada em funções como chave (*key*), consulta (*query*) e valor (*value*). A aplicação desses pesos é demonstrada na Equação 2.7.

$$q_i = W^Q x_i; k_i = W^K x_i; \upsilon_i = W^V x_i, \tag{2.7}$$

onde x_i é a entrada e q_i, k_i e υ_i são os valores após aplicar os pesos de consulta (*query*), chave (*key*) e valor (*value*), respectivamente.

Dado esses cálculos, o valor de escore, anteriormente apresentado na Equação 2.3, agora é obtido pelo produto entre o vetor de consulta (*query*) e o vetor de chaves (*key*), conforme ilustrado na Equação 2.8.

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

onde x_i e x_j são valores do vetor de entradas, enquanto q_i e k_j são os valores após a aplicação dos pesos de consulta (*query*) e chave (*key*), respectivamente.

O cálculo da função de probabilidade, a *softmax*, permanece o mesmo da Equação 2.5, utilizando o novo valor de escore obtido na Equação 2.8. No entanto, a saída é fundamentada nos valores do vetor de valor, conforme apresentado na Equação 2.9.

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

onde y_i é a saída para a correspondente entrada x_i, distribuição de probabilidades α_i e vetor υ de tamanho j correspondentes.

Para evitar perdas de gradiente durante os treinamentos, é feita uma escalabilidade do produto conforme a Equação 2.10.

$$score(x_i,x_j) = \frac{q_i.k_j}{\sqrt{d_k}}, \tag{2.10}$$

onde x_i e x_j são valores do vetor de entradas, q_i e k_j são os valores após a aplicação dos pesos de consulta (*query*) e chave (*key*), respectivamente. Por fim, temos o valor d_k, que representa o tamanho do *embedding*.

O processo explicado pode ser observado na Figura 2.9.

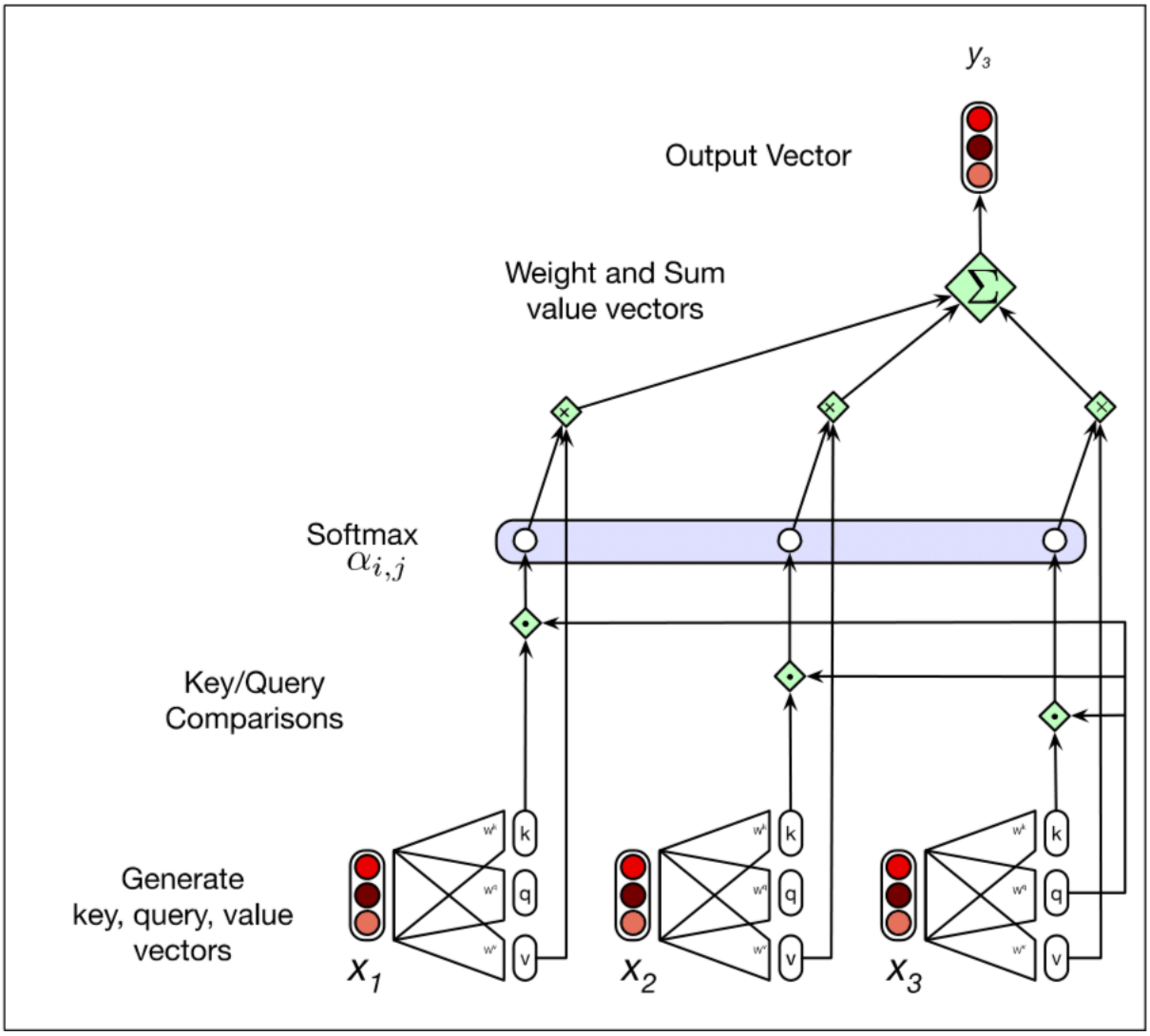


Figura 2.9 – Fluxo de informação da camada *self-attention*.

Fonte: Jurafsky e Martin (2023)

2.5.3 Conexão residual e Normalização

De acordo com Jurafsky e Martin (2023), a camada de *Residual connection* tem o objetivo de repassar informações das camadas inferiores para as superiores, evitando passar pelas camadas intermediárias. Isso permite que as camadas superiores possuam informações sobre os dados iniciais e todas as camadas anteriores. Essa conexão é implementada pela cópia da entrada e pela adição da saída das camadas intermediárias. O resultado é normalizado

pela camada de normalização *self-attention*, representada pela Equação 2.11, e pela camada *feedforward* (FFN), representada pela Equação 2.12.

$$z = LayerNorm(x + Self Attention(x)), \tag{2.11}$$

onde LayerNorm representa a normalização da camada de *self-attention* e z representa a saída e x a entrada da *self-attention*.

$$y = LayerNorm(z + FFN(z)), \tag{2.12}$$

onde LayerNorm representa a normalização da camada de *feedfoward*, enquanto y representa a saída e z a entrada da *feedfoward* .

Segundo Jurafsky e Martin (2023), a normalização realizada na camada de *Normalize*, que visa aprimorar o desempenho de treinamento em redes de aprendizado profundo, é calculada inicialmente obtendo a média através da Equação 2.13 e, em seguida, o desvio padrão dos valores de entrada por meio da Equação 2.14.

$$\mu = \frac{1}{d_h}\sum_{i=1}^{d_h} x_i, \tag{2.13}$$

onde μ representa a média da entrada, x_i é um valor do vetor X de entradas e d_h representa a quantidade de valores no vetor X.

$$\sigma = \frac{1}{d_h}\sum_{i=1}^{d_h}(x_i - \mu)^2, \tag{2.14}$$

onde μ representa a média, σ representa o desvio padrão da entrada, x_i é um dos valores do vetor X de entradas e d_h representa a quantidade de valores no vetor X.

A partir desses valores é feito uma normalização no vetor X com valor igual a média μ de modo a deixar a média em zero. Além disso, o vetor resultante é dividido pelo seu desvio padrão σ de modo a igualar o desvio padrão do vetor resultante $\hat{x}$ a 1 como mostrado na Equação 2.15.

$$\hat{x} = \frac{(x - \mu)}{\sigma}, \tag{2.15}$$

Por fim, é aplicado um valor de ganho γ em conjunto com um deslocamento β de acordo com a Equação 2.16 para efetuar um último ajuste.

$$LayerNorm(x) = \gamma\hat{x} + \beta, \tag{2.16}$$

2.5.4 Feedfoward

Segundo Jurafsky e Martin (2023), a camada de *feedfoward* é de extrema importância na arquitetura do *Transformer*. A camada consiste em uma rede neural geralmente simples e completamente conectada, onde cada componente da rede neural recebe informação de todos os componentes da camada anterior. A Figura 2.10 demonstra a organização da camada.

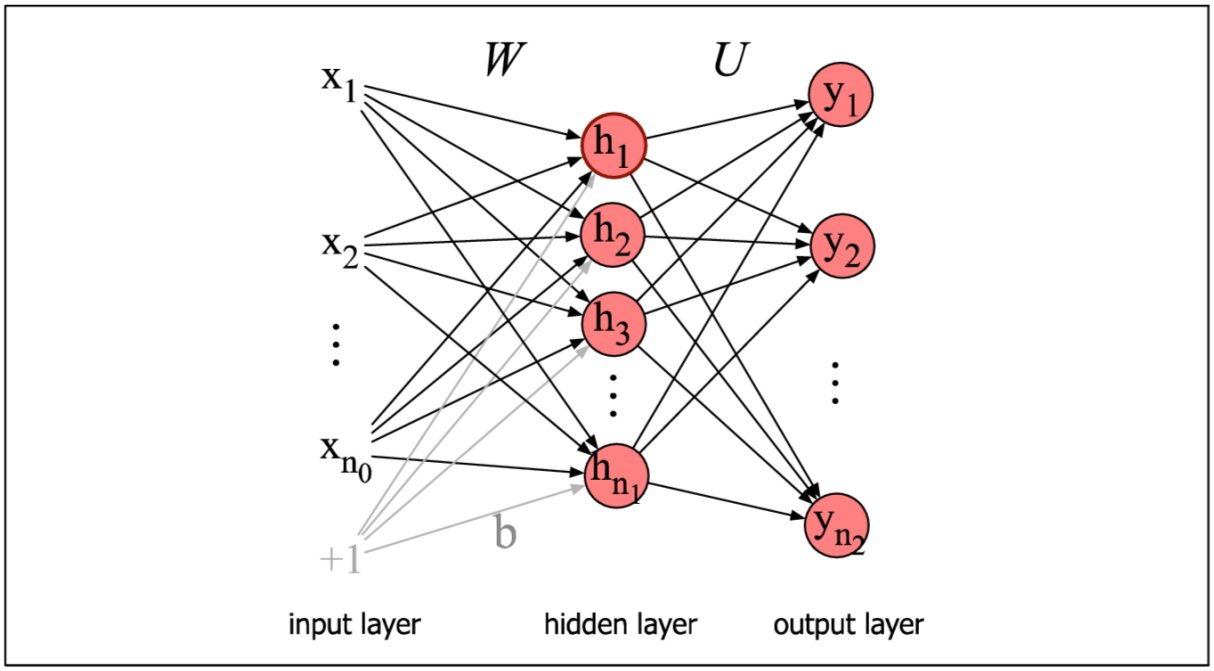


Figura 2.10 – Organização do *Feedfoward*.

Fonte: (JURAFSKY; MARTIN, 2023)

Além de transferir os valores que recebe, a camada atribui um peso para cada uma dessas informações, aplicando não-linearidades nesses resultados, com o objetivo de aumentar a eficiência do treinamento do *Transformer*. Esse processo é demonstrado pela Equação 2.17.

$$h = \sigma(Wx + b), \tag{2.17}$$

onde h é um vetor de saídas de uma camada interior qualquer da rede neural da camada de *feedfoward*, W é uma uma matiz de pesos aplicado no vetor de entradas x adicionado ao vetor de *bias* b e por fim aplicando não-linearidades com a função de ativação, onde neste exemplo foi utilizada a função *sigmoid* representado pelo simbolo σ.

Após passar por todas as camadas internas da rede neural de *feedforward*, é realizado um último ajuste com a matriz de pesos U por meio da seguinte fórmula, que cria o vetor de saída z da camada de *feedforward*, sendo h uma camada camada interior qualquer da rede neural.

$$z = Uh. \tag{2.18}$$

2.6 Métricas de avaliação

No estudo de processamento de linguagem natural é comum a utilização de métricas como **precision**, **recall**, **F-score** e **accuracy** para avaliar os modelos de classificação.

No contexto de classificação binária, as classificações são dadas em apenas dois rótulos, positivo ou negativo, a tabela de contenção é apresentada em uma tabela de quatro células, denominada de matriz de confusão (POWERS, 2008).

A matriz de confusão é uma tabela indicativa dos erros e acertos os comparando com o resultado esperado. Verdadeiro Positivo (*True Positive* - TP) e Falso Positivo (*False Positive* - FP) refere-se aos valores de predição corretos e incorretos, respectivamente e similarmente, Verdadeiro Negativo (*True Negative* - TN) e Falso Negativo (*False Negative*- FN) correspondem aos valores de predição corretos e incorretos, respectivamente.

Tabela 2.1 – Matriz de Confusão.

		Predição	
		Negativo	**Positivo**
Real	**Negativo**	Verdadeiro Negativo	Falso Positivo
	Positivo	Falso Negativo	Verdadeiro Positivo

A partir dos quantitativos mostrados na Tabela 2.1, pode-se calcular a precisão (*precision*), sensibilidade (*recall*), e a acurácia (*accuracy*).

A ***accuracy*** (acurácia) é a métrica de utilizada para avaliação de quantas classificações foram de fato consideradas corretas e é calculada de acordo com a Equação 2.19.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}, \tag{2.19}$$

em que TP = Verdadeiro Positivo, FP = Falso Positivo, TN = Verdadeiro Negativo e FN = Falso Negativo.

A ***precision*** (precisão) é a parcela de classe Positiva correta dentre todas as classificações positivas do modelo:

$$Precision = \frac{TP}{TP + FP}, \tag{2.20}$$

onde TP = Verdadeiro Positivo e FP = Falso Positivo.

Recall, também chamado de *Sensibilidade*, é a proporção de Reais Positivos que o modelo classificou como positivo:

$$Recall = \frac{TP}{TP + FN}, \tag{2.21}$$

onde TP = Verdadeiro Positivo e FN = Falso Negativo.

O ***F1-score*** é uma métrica de teste de acurácia. A medida é dada a partir da média harmônica dos cálculos de precisão e *recall* (WIKIPEDIA CONTRIBUTORS, 2023):

$$F1 - Score = 2 \times \frac{Precision * Recall}{Precision + Recall}. \tag{2.22}$$

2.6.1 Acurácia Top-N

Uma métrica também utilizada para avaliação de tarefas de classificação em *Machine Learning* é a acurácia top-N (*Top-N accuracy*). Nesse método de avaliação, não é levado em conta apenas a classificação correta da classe prevista, e sim se o valor de classificação está entre as N classes. (LEE et al., 2016)

A acurácia top-N é dada pela proporção do número total de casos em que a classe verdadeira está entre as N melhores classificadas pelo o número total de casos.

2.6.2 Equal Error Rate

A *Equal Error Rate* (EER) é utilizada para a avaliação do valor limiar para a curva da taxa de falsa aceitação (*False Reject Rate (FRR)*) e a curva de taxa de falsa rejeição (*False Accept Rate (FAR)*).

O limiar é definido como o valor em que a taxa de falsa rejeição e a taxa de falsa aceitação são iguais, ou seja, o ponto em que a FRR é igual a FAR.

A curva da taxa de falsa aceitação representa o caso em que as entradas inválidas são consideradas aceitas, ou seja, são os falsos positivos, e a taxa de falsa rejeição são os casos em que entradas válidas são erroneamente classificadas como inválidas, ou seja, os falsos negativos.

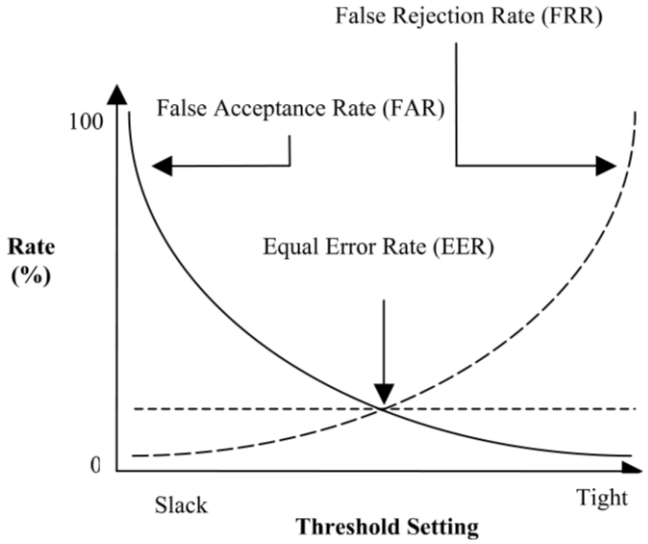


Figura 2.11 – Gráfico para o cálculo da EER.

Fonte: (CLARKE; FURNELL; REYNOLDS, 2002)

A Equal Error Rate (EER) representa o limiar limite de escolha, porque representa o ponto de encontro entre as curvas FAR e FRR, conforme Figura 2.11. O valor ótimo é muito próximo de zero.

Após estabelecer a fundamentação teórica referencial para o trabalho e detalhar as métricas de avaliação dos resultados esperados, o Capítulo 3 abordará os trabalhos correlacionados que serviram como base para o estudo e apresentar alguns modelos de *Transformers* utilizados.

3 Trabalhos relacionados

Os contratos de licitações de obras públicas movimentam bilhões de reais anualmente, podendo ser então uma facilitação para a fraude no processo. Para evitar conluio, superfaturamento e fraudes nesse contexto, têm sido empregados esforços para a detecção do ato ilegal. Em Lima et al. (2020) foi proposto a utilização de técnicas de Processamento de Linguagem Natural (NLP) para a classificação de conluio. O trabalho utilizou 15.132.968 entradas textuais provenientes de licitações públicas com 1.907 sendo entradas com possibilidade de serem fraudes. O DNN (*Deep Neural Network*) baseado em LSTM (*Long Short-Term Memory*)[1] aplicado foi comparado com diversos *baselines* e classificadores clássicos e obtiveram melhores resultados.

No Brasil são publicadas informações de interesse público, como as licitações públicas e a contratação de empresas em seus Diários Oficiais. Os altos volumes de publicações é de difícil acesso se tornando custoso a dimensão de informações relevantes com características governamentais, como a efetividade de políticas públicas ou a existência de sistemas ilegais. O trabalho de Carvalho et al. (2022) propõe a plataforma *Deep-Vacuity* que coleta dados provenientes dos Diários Oficiais e a partir de técnicas de Aprendizado de Máquina consegue coletar, depurar, analisar e consolidar dados por meio de uma interface gráfica, de maneira a auxiliar tomadores de decisão no combate a corrupção.

3.1 Uso de Modelos com *Transformers*

O *Sentence-BERT* (SBERT) foi publicado em Reimers e Gurevych (2019), onde é apresentada uma modificação do modelo BERT pré-treinado para encontrar as sentenças semanticamente similares com a utilização da similaridade de cosseno. Quando comparado com os modelos BERT (*Bidirectional Encoder Representations for Transformers*) e RoBERTa (*Robustly Optimized BERT Pretraining Approach*), apresentados anteriormente, o esforço reduz drasticamente.

Em Sanh et al. (2019) é apresentado um método de pré-treinamento para o modelo **DistilBERT**. O trabalho utilizou o uso de destilação para a construção de modelos específicos e foi capaz de reduzir o tamanho de um modelo BERT em 70%, mantendo a capacidade de compreensão em 97% sendo 60% mais rápido.

No trabalho apresentado por Wenhui Wang et al. (2020b) é relatada uma nova abordagem para compactar *Transformers* baseados em modelos pré-treinados. A ideia do artigo

[1] LSTM (*Long Short-Term Memory*) é uma arquitetura de rede neural recorrente (RNN) utilizada em aprendizado de sequência de dados. As aplicações comuns de LSTM incluem as tarefas de processamento de linguagem natural. (MATHWORKS, 2023)

é imitar os módulos de *self-attention* em modelos professor e aluno. O trabalho propõe a separação do módulo de *self-attention* da última camada do *Transformers* do modelo professor, além de introduzir o produto escalar entre os valores de *self-attention*. O estudo traz também a introdução de um professor-assistente aumentando o desempenho do modelo.

Os métodos apresentados são modelos geralmente voltados para uma linguagem. Pensando em expandir os modelos existentes para outras línguas, Reimers e Gurevych (2020) propôs um novo treinamento para sentenças traduzidas. A partir do modelo original, é gerado os *embeddings* na nova língua e então é realizado o treinamento de maneira a reproduzir o comportamento do modelo original. O trabalho demonstrou efetividade em mais de 50 línguas.

No trabalho apresentado em Souza, Rodrigo Nogueira e Lotufo (2020a) foi estudada a possibilidade de utilização das capacidades de modelos de *Transformers* para tarefas de NLP em português. O modelo foi avaliado nas tarefas de Similaridade Semântica (STS), Reconhecimento de Entidades Nomeadas (NER) e Inferência Textual (RTE).

A tabela a seguir mostra métricas de performance de alguns modelos baseados nos trabalhos definidos anteriormente.

Tabela 3.2 – Comparação modelos *sentence-transformers*.

Modelo	Performance Média	Velocidade	Conjunto de Dados Treinamento	Tamanho
mpnet (TRANSFORMERS, 2023d)	63,30	2800	1B+	420 MB
qa mpnet (TRANSFORMERS, 2023h)	62,18	2800	215M	420 MB
DistilRoBERTa (TRANSFORMERS, 2023a)	59,84	4000	1B+	290 MB
MiniLM L12 (TRANSFORMERS, 2023b)	59,76	7500	1B+	120 MB
qa DistilBERT (TRANSFORMERS, 2023f)	59,41	4000	215M	250 MB
MiniLM L6 (TRANSFORMERS, 2023c)	58,80	14200	1B+	80 MB
qa MiniLM L6 (TRANSFORMERS, 2023g)	58,08	14200	215+	80 MB

3.2 Similaridade Semântica

O artigo de Resnik (1995), demonstra alguns dos principais problemas de modelos de similaridade semântica como ambiguidades, demonstrando que em certos contextos palavras que deveriam ser similares podem se mostrar muito diferentes semanticamente. O artigo então propõe um modelo de similaridade feito por meio de um modelo de taxonomia *IS-A* e demonstrou resultados consideravelmente superiores a modelos tradicionais na época, utilizando a noção de conteúdo da informação, considerando aspectos polissêmicos das palavras.

Já o trabalho realizado por Slimani (2013) faz uma análise dos diversos modelos de similaridade semântica em diversas categorias como estrutura, conteúdo da informação e características das técnicas. O trabalho continua com uma avaliação crítica dos diferentes

métodos baseados em dois métodos de referência, tentando permitir que outros usuários de similaridade semântica consigam escolherem o método ideal para cumprir os seus requisitos.

O trabalho de Majumder et al. (2016) apresenta uma análise de diferentes métodos de similaridade semântica comparando as diversas vantagens e desvantagens dos mais relevantes, suas principais utilizações e identificando três principais categorias dos métodos (i) *Topological/Knowledge-based* (ii) *Statistical/Corpus Based* (iii) *String based*, além de propor um modelo chamado *WordNet*, que é um método Hibrido entre as categorias em que foram colocados os outros modelos, e obteve ótimos resultados.

O trabalho apresentado por Fonseca et al. (2016), traz resultados da avaliação de duas tarefas de processamento de linguagem naturais que tratam de pares de trechos. A primeira tarefa é a Inferência Textual, em que o objetivo é determinar se o significado de um trecho implica no segundo. Já a segunda tarefa é a Similaridade Semântica através de uma pontuação. O corpus utilizado no trabalho foi composto de sentenças extraídas de textos jornalísticos na língua portuguesa, nas variantes europeia e brasileira. A pesquisa foi separada em equipes com diferentes estratégias abordada para as tarefas como redes semânticas, modelos de espaço vetorial, redes neurais.

4 Metodologia proposta

Para estabelecer o relacionamento semântico entre publicações do DOU por meio dos modelos de *Transformers* o trabalho foi dividido em etapas, conforme a Figura 4.12. A seguir serão especificadas detalhadamente cada etapa de projeto proposta.

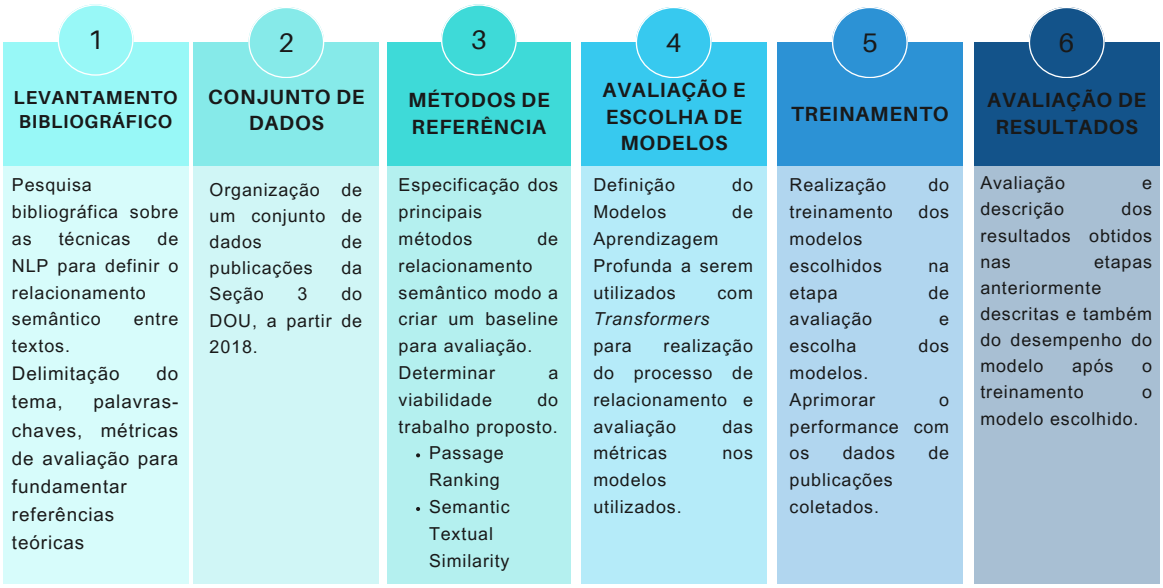


Figura 4.12 – Metodologia proposta.

Conforme a Figura 4.12, a metodologia proposta é realizada de maneira sequencial, e a execução de uma etapa depende da finalização da anterior. É importante ressaltar que a execução de cada passo é de imprescindível realização para atingir os objetivos descritos no Capítulo 1.

4.1 Levantamento Bibliográfico

O ponto de partida do trabalho é dado pelo levantamento bibliográfico, onde a partir da definição de alguns critérios principais será definido os trabalhos correlacionados apresentados no Capítulo 3.

Durante esta etapa, os pontos fundamentais incluem a definição e delimitação do tema, a escolha de palavras-chave, a avaliação dos resultados obtidos nos documentos encontrados e a análise de sua relevância para o presente estudo, de maneira a fundamentar as referências teóricas a serem utilizadas.

O sítio eletrônico Connected Papers é de grande valia durante a etapa de coleta bibliográfica, pois auxiliou na busca dos trabalhos relacionados demostrando a relevância do publicado anteriores e após com o assunto do presente relatório.

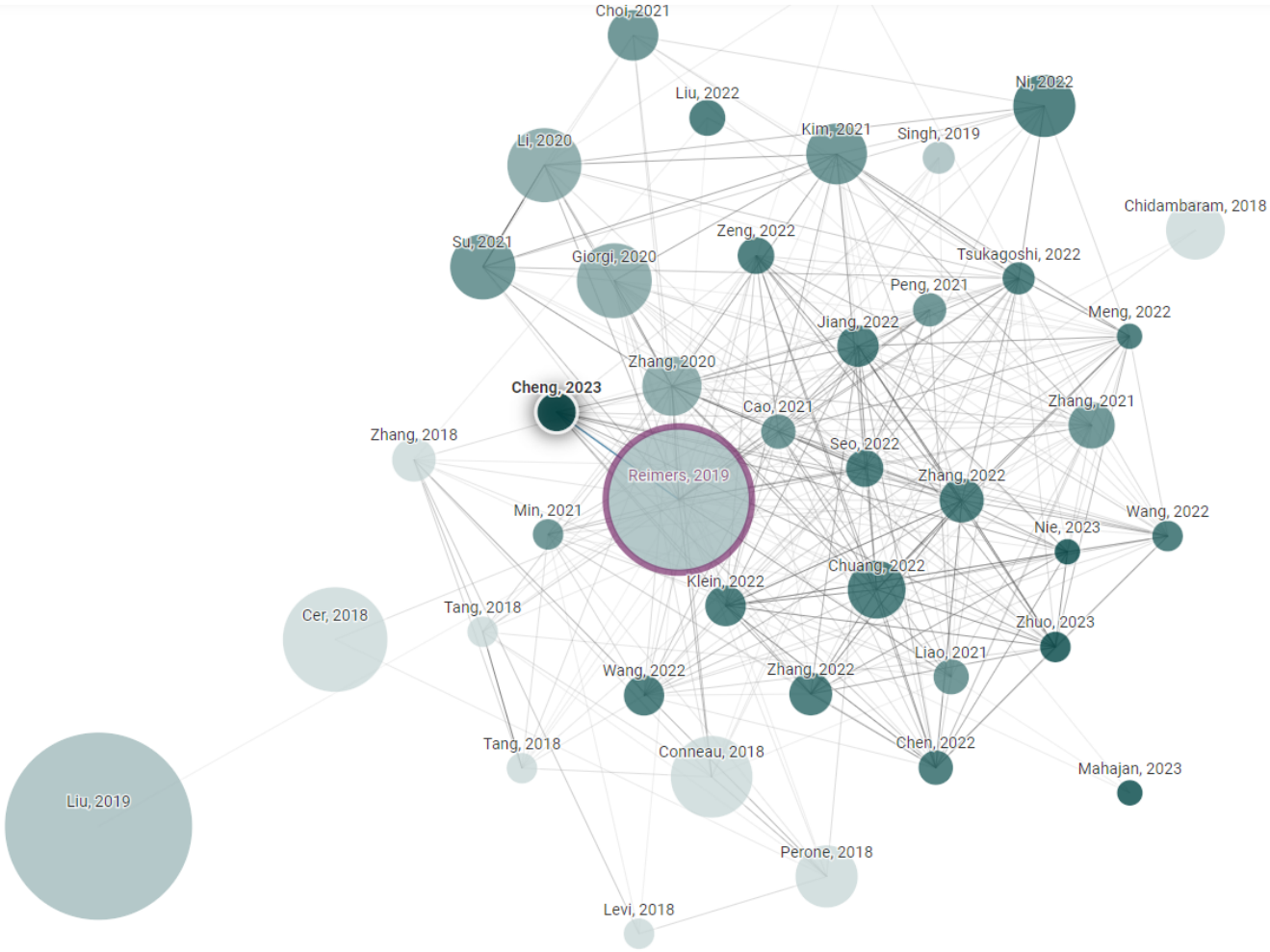


Figura 4.13 – Connected Papers.

Fonte: (CONNECTED PAPERS, 2023)

4.2 Conjunto de Dados

O conjunto de dados é proveniente de publicações da Seção 3 do Diário Oficial da União. Foram coletadas publicações a partir de 2018, pois as publicações anteriores a janeiro de 2018 estão disponíveis apenas nas versões certificadas, em formato PDF, no diário oficial completo sem a separação de publicações. Toda a separação de publicações foi realizada de maneira manual, sendo analisadas cada publicação e seu conjunto de convênios um a um.

As publicações coletadas foram referentes a Extratos de Convênios, Extratos de Termo Aditivo, Extratos de Prorrogação de Ofício, Avisos de Licitações e Extratos de Contratos e seus respectivos Termos Aditivos. A correlação a ser estabelecida entre esses elementos é realizada conforme Figura 2.1.

A partir da coleta das publicações diretamente do DOU, serão anotadas manualmente quatro informações principais de cada publicação:

1. Convênio - convênio em que a publicação está vinculado;
2. Id - identidade única de cada publicação da lista de convênio;
3. URL - texto com link para publicação dentro do site do DOU; e
4. Relação - especificação da vinculação entre as publicações.

A partir de técnicas de extração de dado de modo automático por meio de requisições HTTP serão obtidas de cada publicação o texto de descrição. Na Figura 4.14 é indicado em vermelho o texto que será extraído de cada uma das publicações coletadas. A extração será realizada pela a retirada do conteúdo da classe *"douparagraph"* do sítio do Diário Oficial da União.

DIÁRIO OFICIAL DA UNIÃO

Publicado em: 06/09/2018 | Edição: 173 | Seção: 3 | Página: 46

Órgão: Ministério da Defesa/Secretaria-Geral

EXTRATO DE CONVÊNIO

Espécie: Convênio Nº 864119/2018, Nº Processo: 60414000184201813, Concedente: MINISTERIO DA DEFESA, Convenente: MUNICIPIO DE JI-PARANA CNPJ nº 04092672000125, Objeto: Pavimentação em vias urbanas com drenagem, meio-fio, sarjetas e calçadas., Valor Total: R$ 346.800,00, Valor de Contrapartida: R$ 6.800,00, Valor a ser transferido ou descentralizado por exercício: 2018 - R$ 340.000,00, Crédito Orçamentário: Num Empenho: 2018NE800192, Valor: R$ 340.000,00, PTRES: 140258, Fonte Recurso: 0100000000, ND: 44425141, Vigência: 05/09/2018 a 20/08/2021, Data de Assinatura: 05/09/2018, Signatários: Concedente: ROBERTO DE MEDEIROS DANTAS CPF nº 483.922.198-72, Convenente: JESUALDO PIRES FERREIRA JUNIOR CPF nº 042.321.878-63.

Este conteúdo não substitui o publicado na versão certificada.

Figura 4.14 – Descrição da publicação.

Fonte: (IMPRENSA NACIONAL, 2023)

4.3 Métodos de Referência

A partir das referências bibliográficas, os métodos de referência serão elencados para definir os melhores resultados a serem utilizados como forma de comparação com os modelos de *Transformers*. É importante ressaltar que as implementações iniciais também são relevantes para determinar a viabilidade do trabalho proposto. Sem essa etapa, não será possível indicar se os *Transformers* são adequados para a tarefa.

Nesse estágio, os métodos de referências a serem utilizados são provenientes da HuggingFace[1]. O objetivo é avaliar se uma publicação faz parte de um determinado convênio, e para isso, é preciso extrair as informações dos documentos e então determinar pelos modelos se aquela publicação é similar ao convênio de busca.

Dessa forma, serão avaliados os modelos de *Passage Ranking* (NOGUEIRA, R. F.; JIANG; LIN, 2020) e os modelos de *Semantic Textual Similarity* (CER et al., 2017). A tarefa de *Passage Ranking* consiste em ordenar documentos de acordo com sua relevância para uma consulta, utilizando a métrica MRR - *Mean Reciprocal Rank*. Os modelos de *Passage Ranking* recebem diversos documentos e uma referência de busca, realizando a classificação

[1] Hugging Face é uma empresa que se concentra no desenvolvimento de tecnologias relacionadas ao processamento de linguagem natural (NLP) e aprendizado de máquina. A empresa mantêm a biblioteca de *Transformers*, que contém implementações de modelos de linguagem pré-treinados, como o GPT (*Generative Pre-trained Transformer*) e BERT.

dos documentos com base em sua similaridade com a referência, e retornam os escores do ranqueamento. Já a tarefa de *Semantic Textual Similarity* tem como objetivo determinar quão similares são os textos em termos de significado. Nessa tarefa, a partir de um texto de referência, uma lista de textos é avaliada para medir a similaridade entre eles, resultando em uma lista de escores.

4.4 Avaliação e Escolha de Modelos

Após a definição dos modelos de referência e os valores de escores obtidos para a comparação semântica entre as publicações do DOU, será possível realizar a avaliação desses modelos de acordo com as métricas analisadas na Seção 2.6.

Para cada conjunto de convênios, será efetuado o cálculo da média dos escores de similaridade semântica entre as publicações, a fim de realizar a comparação de ranqueamento. Esta comparação não ocorrerá entre publicação, mas sim em grupos de publicações derivadas de um mesmo convênio. Com base nos valores calculados, será obtida as porcentagens de convênios classificados corretamente, seguindo os critérios do top 1 e do top 5, conforme descrito na Seção 2.6.1.

Após obter os valores de ranqueamento para os conjuntos de convênios, a matriz de confusão, precisão, acurácia, *recall* e a *equal error rate* serão também calculadas, porém com os escores individuais por publicação.

É esperado que ao calcular todas as métricas para todo o conjunto de dados nos métodos de referência, seja estendível para outros modelos, tanto em português, como em inglês. Os modelos com os melhores resultados serão então treinados.

4.5 Treinamento

A partir da modelagem e da construção do conjunto de dados, por fim, foi a última etapa consiste no treinamento dos modelos de *Transformers* pré-treinados. O objetivo do treinamento é aprimorar a performance do modelo treinado, utilizando os dados de publicações coletadas da Seção 3 do DOU.

4.5.1 Conjunto de dados do treinamento

Os modelos serão treinados em conjuntos de dados que contêm pares de textos associados a um escore. Esse escore é definido como 0 ou 1, onde publicações de um mesmo conjunto de convênios é atribuído 1, ou seja, deve ter o maior nível de similidade semântica com a publicação do extrato de convênio e publicações de convênios distintos é atribuído 0, isto é, não devem indicar similidade semântica.

```
[
  {
    "texto": "ESPÉCIE: 01 Prorroga de Ofício - Convênio nº 882007/2018, MINISTÉRIO DA CIDADANIA, Unidade Gestora: 1800073 - Gestão: 00001; Prefeitura Municipal de Pompéu/MG, CNPJ: 18.296.681/0001-42 - P.I. 424/2016, Art. 27, VI e DECRETO Nº 10.315, DE 6 DE ABRIL DE 2020. Vigência: 28/12/2018 a 31/12/2020. Data de Assinatura: 22/04/2020. Assina: MINISTÉRIO DA CIDADANIA - FABIOLA PULGA MOLINA - Secretaria Nacional de Esporte, Educação, Lazer e Inclusão Social. Processo nº 58000.003259/2018-20.",
    "convenio_comparado": "Espécie: Convênio 882007/2018, Nº Processo: 58000.003259/2018-20, Concedente: MINISTÉRIO DO ESPORTE, por meio da SECRETARIA NACIONAL DE ESPORTE, EDUCAÇÃO, LAZER E INCLUSÃO SOCIAL. Convenente: MUNICÍPIO DE POMPÉU/MG. CNPJ nº 18.296.681/0001-42, Objeto: 'Realização do Festival Mexa-se Superação, no Município de Pompéu/MG', conforme detalhado no Plano de Trabalho, Valor Total: R$ 102.832,25, Valor de Contrapartida: R$ 2.832,25, Valor a ser transferido ou descentralizado por exercício: 2018 - R$ 100.000,00, Crédito Orçamentário: Num Empenho: 2018NE801166, Valor: R$ 100.000,00, PTRES: 142085, Fonte Recurso: 0100, ND: 334041, Vigência: 28/12/2018 a 28/12/2019, Data de Assinatura: 28/12/2018, Signatários: Concedente: ANGELO DE BORTOLI FILHO CPF nº 106.987.118-40, Convenente: OZÉAS DA SILVA CAMPOS CPF nº 008.438.166-35._",
    "valor_esperado": 1
  },
  {
    "texto": "Terceiro Termo Aditivo ao Contrato nº 021/2018 - Tomada de Preço nº 002/2018, Convênio OGU CR 829584/2016 - Processo nº 1029613-96/206, firmado entre o Município de Moema e o Ministério das Cidades, recapeamento asfáltico de vias públicas. Objeto: Alteração do valor global do Contrato, em conformidade com a Planilha Orçamentária apresentada na reprogramação do Convênio. Contratada: Empreser - Empresa de Prestação de Serviços LTDA, CNPJ nº 19.268.374/0001-10. Data da assinatura: 14/06/2019. Documento completo no site: www.moema.mg.gov.br, aba Diário Oficial._",
    "convenio_comparado": "Espécie: Convênio 882007/2018, Nº Processo: 58000.003259/2018-20, Concedente: MINISTÉRIO DO ESPORTE, por meio da SECRETARIA NACIONAL DE ESPORTE, EDUCAÇÃO, LAZER E INCLUSÃO SOCIAL. Convenente: MUNICÍPIO DE POMPÉU/MG. CNPJ nº 18.296.681/0001-42, Objeto: 'Realização do Festival Mexa-se Superação, no Município de Pompéu/MG', conforme detalhado no Plano de Trabalho, Valor Total: R$ 102.832,25, Valor de Contrapartida: R$ 2.832,25, Valor a ser transferido ou descentralizado por exercício: 2018 - R$ 100.000,00, Crédito Orçamentário: Num Empenho: 2018NE801166, Valor: R$ 100.000,00, PTRES: 142085, Fonte Recurso: 0100, ND: 334041, Vigência: 28/12/2018 a 28/12/2019, Data de Assinatura: 28/12/2018, Signatários: Concedente: ANGELO DE BORTOLI FILHO CPF nº 106.987.118-40, Convenente: OZÉAS DA SILVA CAMPOS CPF nº 008.438.166-35.",
    "valor_esperado": 0
  }
]
```


Figura 4.15 – Conjunto de dados para o treinamento.

4.5.2 Validação Cruzada

A validação cruzada (*cross-validation*) é uma técnica utilizada para avaliar a capacidade de um modelo de generalizar para novos dados e ajuda a estimar o desempenho de um modelo de forma mais robusta do que apenas dividir os dados em conjuntos de treinamento e teste. (MICROSOFT, 2023)

A ideia da validação cruzada é dividir o conjunto de dados em várias partições (chamadas de "*fold*"), geralmente K partições. Em seguida, o modelo é treinado e avaliado K vezes, onde em cada iteração, uma das partições é usada como conjunto de teste e as outras K-1 partições são usadas como conjunto de treinamento. Isso permite que cada parte do conjunto de dados seja usada tanto para treinamento quanto para teste em iterações diferentes. (BROWNE, 2000)

Segundo Microsoft (2023), o método de avaliação cruzada segue as seguintes etapas:

- Divisão dos dados em K partições iguais (folds) de maneira aleatória;
- Treinamento e teste, onde o modelo é treinado K vezes. Em cada iteração, uma partição é utilizada como conjunto de teste e as outras K-1 partições do conjunto de dados é utilizada para treinamento; e
- Avaliação dos resultados das K iterações, onde as estatísticas de precisão são usadas como estimativa do desempenho do modelo.

A Figura 4.16 demonstra esse processo de validação cruzada e a maneira como é realizada a separação dos dados em teste e treino a cada iteração.

Figura 4.16 – Processo Validação Cruzada.

Fonte: Scikit Learn (2023)

4.5.3 Função de perda

A função de perda (*loss function*) desempenha um papel crucial quando realizado o treinamento de um modelo pré-treinado, porque determina como o modelo de *embeddings* funcionará para a tarefa realizada. (REIMERS; GUREVYCH, 2019)

Para cada tarefa e conjunto de dados do treinamento é preciso avaliar qual função de perda se encaixa no objetivo de trabalho. Segundo Reimers e Gurevych (2019), algumas das funções que podem ser utilizadas no treinamento de *sentence-transformers* são:

- *BatchAllTripletLoss*;
- *ContrastiveLoss*;
- *CosineSimilarityLoss*;
- *SoftmaxLoss*; e
- *TripletLoss*.

O conjunto de dados utilizados no treinamento é composto por um par de sentenças com uma pontuação de similaridade entre eles. Dessa forma, é necessário a utilização de funções de perda que otimiza a pontuação das sentenças com uma pontuação mais próxima no espaço vetorial, dessa forma a função de perda a ser utilizada é a **CosineSimilarityLoss**.

A similaridade de cossenos, conforme apresentado na Seção 2.4.1, é calculado por meio dos *embeddings* das sentenças. A partir da pontuação do conjunto de dados do treinamento, o modelo é ajustado para reconhecer as semelhanças das sentenças a serem avaliadas (REIMERS; GUREVYCH, 2019).

4.5.4 Avaliação do Treinamento

Após o ajuste do conjunto de dados de treinamento e a definição da função de perda é possível realizar o processo de aprimoramento dos modelos de *Sentence Transformer* definidos na Seção4.4.

A avaliação do treinamento será feita em duas etapas, onde na primeira etapa será avaliado os resultados do treinamento realizado por meio da validação cruzada e a segunda etapa consiste na comparação entre o modelo treinado com o modelo pré-treinado.

Após o treinamento, será realizada a avaliação dos resultados obtidos na aplicação da metodologia apresentada no Capítulo 4 por meio das métricas definidas na Seção 2. Os resultados serão consolidados no Capítulo seguinte.

4.6 Avaliação dos Resultados

Por fim, na etapa de avaliação dos resultados, será feita uma compilação e discussão dos resultados obtidos nas etapas anteriores descritas ao longo do Capítulo 4. As análises abordarão os resultados obtidos na avaliação e escolha de modelos e do treinamento do modelo escolhido por meio da avaliação das métricas de desempenho como *precision*, *recall*, *acurácia*, e também como a estrutura do conjunto de dados montado influenciaram os resultados e os desafios encontrados no processo. Esse ponto do trabalho é de suma importância para concluir os objetivos destacados na Seção 1.3 e contribuir para o estudo da utilização de *Transformers* para encontrar o relacionamento semântico entre publicações da Seção 3 do DOU.

5 Resultados

Neste Capítulo, serão apresentados todos os resultados alcançados na análise de técnicas de obtenção de similaridade semântica utilizando *Transformers*. Os resultados serão organizados de acordo com a metodologia proposta no Capítulo 4.

5.1 Levantamento Bibliográfico

Na primeira etapa de levantamento bibliográfico, apresenta-se os principais resultados alcançados ao explorar a literatura correlacionada ao tema como por exemplo os valores da Tabela 3.2. Primeiramente, obteve-se os trabalhos similares que lidam com a conclusão de conluio em publicações do Diário Oficial da União.

Além disso, também foram alvo de estudos a literatura relacionada ao uso de modelos baseados em *Transformers* para a realização de tarefas de processamento de linguagem natural. Outra análise realizada foi a obtenção de trabalhos que se concentram na obtenção de similaridade semântica textual.

Após a separação da literatura relacionada, foi possível obter os trabalhos especificados no Capítulo 3. Nesse estágio, também foi delimitado o escopo do trabalho, as palavras-chaves e as métricas de avaliação a serem utilizadas.

5.2 Conjunto de Dados

Com os dados coletados e anotados manualmente a partir da seção 3 do DOU a utilização das técnicas de extração de dados em sítios web descritas na Seção 4.2 foi possível a obtenção das publicações no formato de texto. Essas publicações foram posteriormente convertidas para o formato JSON e são do período de janeiro de 2018 a setembro de 2023. A estrutura dos dados segue a estrutura de objeto JSON, conforme descrito pela Figura 5.17

```
{
  "convenio": 8641192018,
  "id": 2,
  "relationship": 0,
  "description": "_CONTRATO N. 096/2020_Processo
    Administrativo: 1-2716/2020, SEMOSP/SEMPLAN oriundo
    do Convênio n. 037/DPCN/2018 - Ministério da Defesa
    - Termo de Convênio SICONV n. 864119/2018.
    Contratada: CAMPEÃO CONSTRUTORA EIRELI - EP.
    Objeto: Pavimentação de vias urbanas com drenagem,
    meio fio, sarjeta e calçadas. Modalidade: Tomada de
    preços n.010/CPL/2020. Prazo: Prazo de vigência do
    contrato e de execução da obra será de 150 (cento e
    cinquenta) dias corridos, conforme cronograma
    físico-financeiro, contados a partir da data do
    recebimento da Ordem de Serviço. Recursos
    Orçamentários: Empenho GL - GLOBAL n. 6672 e 6673.
    Valor: R$ 286.643,50. Foro: Comarca de Ji-
    Paraná/RO.__"
}
```


Figura 5.17 – Estrutura do conjunto de dados em JSON.

A Tabela 5.3 apresenta a definição da estrutura do objeto JSON, incluindo o tipo de dado e uma descrição que esclarece o significado de cada nome dentro do objeto.

Tabela 5.3 – Descrição do conjunto de dados.

Nome	Tipo	Descrição
convenio	long int	Número do convênio que a publicação pertence
id	int	ID da publicação no conjunto de convênio
relationship	int	ID da publicação que o valor está diretamente correlacionado
description	string	Texto da publicação

A partir da extração dos dados no formato JSON, foi possível obter algumas estatísticas referentes ao conjunto de dados, apresentadas na Tabela 5.4.

Tabela 5.4 – Estatísticas do conjunto de dados.

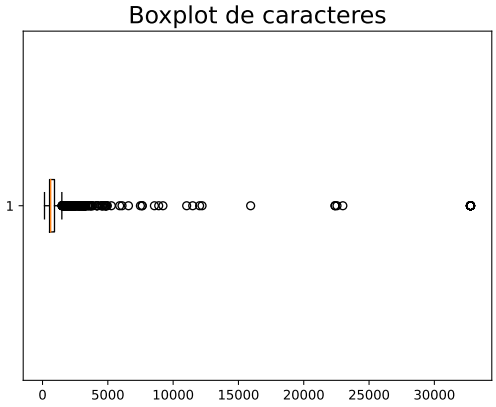
	Dados
Quantidade	1.806 publicações
Máximo	4.735 palavras 32.767 caracteres
Mínimo	16 palavras 151 caracteres
Média	187,43 palavras 1.406,52 caracteres
Desvio padrão	502,92 palavras 3.919,82 caracteres

Foi também conduzida uma análise estatística por conjunto de publicações relacionadas a cada convênio. Os resultados dessa análise estão apresentados na Tabela 5.5.

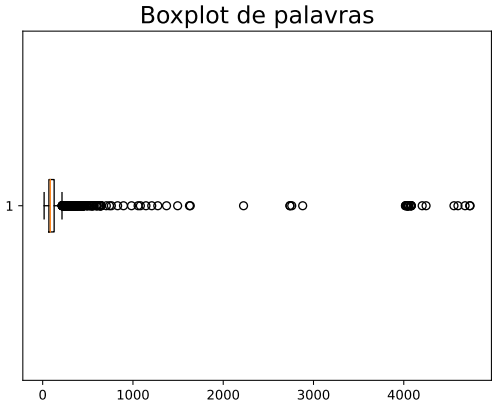
Tabela 5.5 – Estatísticas do conjunto de convênios do conjunto de dados.

	Dados
Quantidade	222 convênios
Máximo	84 publicações
Mínimo	1 publicações
Média	7,13 publicações
Desvio padrão	8,89 publicações

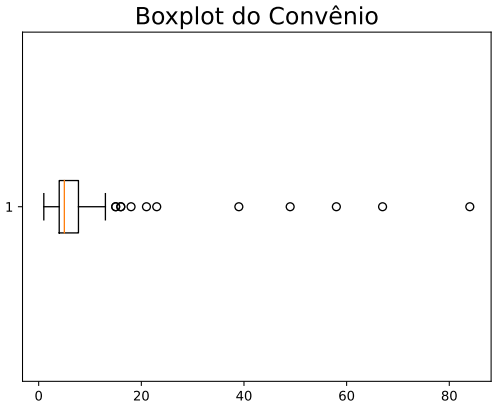
Outra ferramenta utilizada para descrever o conjunto de dados estruturado é o *boxplot*. O *boxplot*, também conhecido como gráfico de caixa, é um gráfico estatístico que é usado para representar a distribuição de um conjunto de dados, permitindo uma representação visual de algumas estatísticas dos dados, como a mediana, o primeiro e terceiro quartil, valores mínimos e máximos e os valores atípicos, chamados de *outliers* (CAPELA, M. V.; CAPELA, J. M., 2011).



(a) Boxplot - Caracteres.



(b) Boxplot - Palavras.



(c) Boxplot - Conjunto de Convênios.

Figura 5.18 – Representação em Boxplot da quantidade de (a) caracteres , (b) palavras e (c) conjunto de convênio.

De acordo com a Figura 5.18, pode-se observar uma dispersão do conjunto de dados. A Figura 5.18a mostra que a mediana é de número de caracteres por publicação é de 638 e que há publicações de mais de 30000 caracteres. A Figura 5.18b referencia que o número mediano de palavras em uma publicação é de 86, mas que existem publicações com mais de 4000 palavras. Por fim, a Figura 5.18c mostra uma menor dispersão em relação ao conjuntos dos convênios. A mediana para a quantidade de publicações referentes a um convênio é de 5 publicações e o valor máximo é de 80.

Apesar do desbalanceamento do conjunto de dados, todos os ou*outliers* foram mantidos para a avaliação do trabalho proposto. Essa escolha foi realizada para avaliar o comportamento da tarefa de similaridade semântica diante de um conjunto despadronizado, conforme as publicações do DOU.

5.3 Métodos de Referência

Após a definição dos modelos a serem utilizados no primeiro momento conforme a Seção 4.3, foi realizado os primeiros testes com 559 dados provenientes da Seção 3 do Diário Oficial. Vale ressaltar a base de dados com 559 publicações é referente a uma partição do conjunto de dados total, ou seja, foram selecionadas aleatoriamente essas publicações para a definição dos métodos de referência. A Figura 5.19 demonstra como foi executada a comparação e determinação da similaridade entre os textos de entrada e a saída após o modelo.

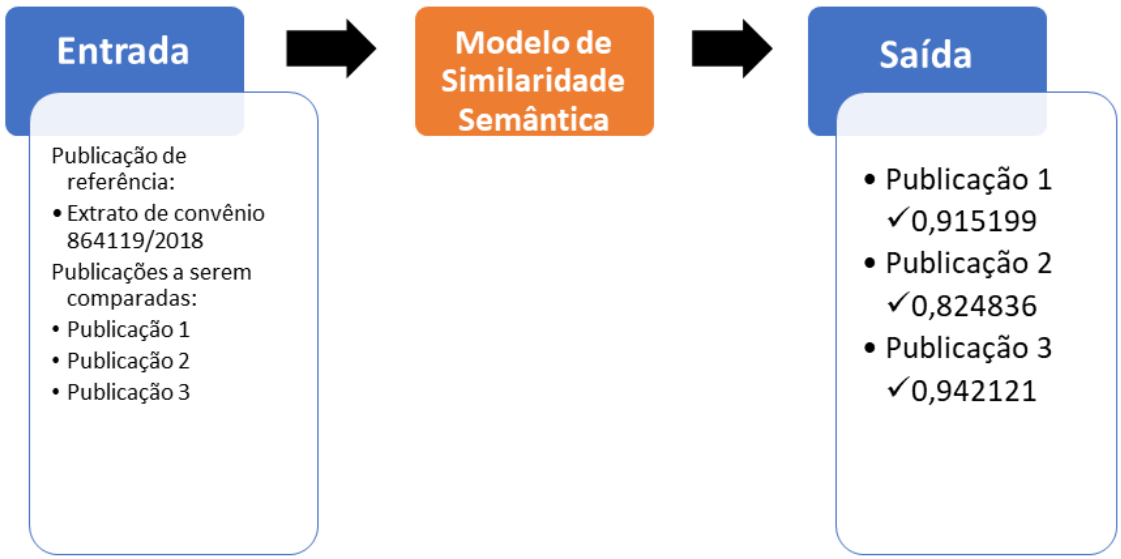


Figura 5.19 – Funcionamento dos modelos de similaridade semântica.

A partir da descrição de um extrato de convênio o modelo de similaridade semântica

vai realizar a comparação dos *embeddings* da publicação de extrato de convênio com os *embeddings* de todas as outras publicações dos conjuntos apresentados. No exemplo apresentado na Figura 5.19, as publicações 1, 2 e 3 obtiveram o valor de similaridade avaliado em 0,92, 0,82 e 0,94, respectivamente. Dessa forma, no exemplo, a publicação 3 é a que possui um maior grau de similaridade semântica.

Para a avaliação da possibilidade da utilização de *Transformers* para a tarefa de determinar a similaridade, foram escolhidos dois modelos iniciais sendo eles:

- *Passage Ranking* (HOFSTÄTTER et al., 2021)
- *Semantic Textual Similarity* (WANG, W. et al., 2020a)

Utilizando os modelos iniciais previamente definidos, em um subconjunto de dados composto por 559 publicações, as métricas de avaliação mencionadas na Seção 2.6 foram calculadas. Os resultados estão apresentados na Tabela 5.7.

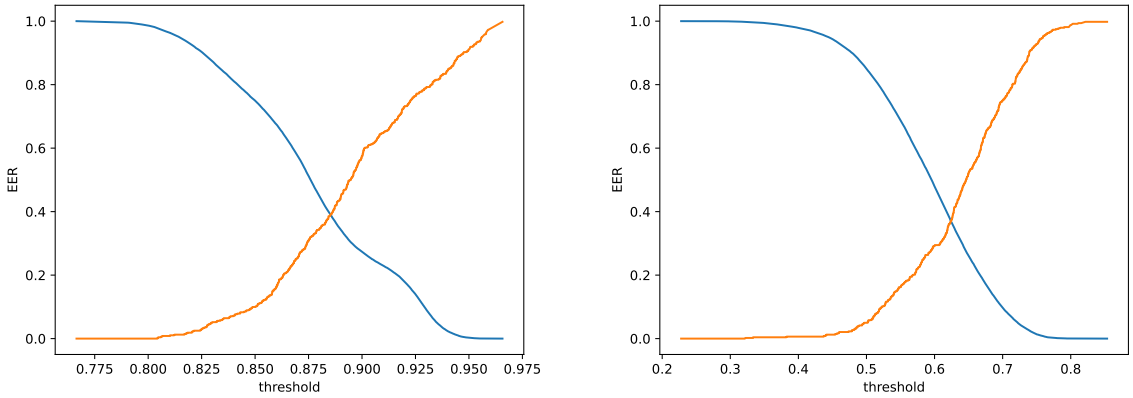
Tabela 5.6 – Acurácia Top-N para modelos de referência

Modelo	**Top 1**	**Top 5**
Passage Ranking	8,97%	16,67%
Semantic Textual Similarity	10,26%	33,33%

Tabela 5.7 – Resultados com a base de 559 publicações.

Modelo	**EER (equal error rate)**	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Passage Ranking	0,39004	0,01991	0,60996	0,03855	0,60999
Semantic Textual Similarity	0,36954	0,02160	0,62863	0,04177	0,63020

Conforme explicitado na Seção 2.6.2, a curva de *equal error rate (EER)* é utilizada para definir a identificação do sistema de acordo com a aceitação e rejeição. Dessa forma, conforme mostrada pela a Figura 2.11, foi realizado a curva de EER para os modelos indicados na Tabela 5.7. Os resultados podem ser vistos na Figura 5.20



(a) EER - *Passage Ranking*.

(b) EER - *Semantic Textual Similarity*.

Figura 5.20 – Gráficos de EER para a (a) *Passage Ranking* e (b) *Semantic Textual Similarity*.

Em resumo, com esses resultados em mão, conclui-se que é possível a obtenção de similaridade semântica para as publicações da Seção 3 do Diário Oficial da União, confirmando a veracidade da metodologia proposta. Também foi possível verificar a avaliação das tarefas de similaridade semântica por meio das tarefas propostas de ranqueamento e de avaliação de similaridade semântica.

5.4 Análise e Estudo de Modelos

Após a verificação da viabilidade da aplicação de modelos de *Transformers* na extração de similaridade semântica textual, conforme definido na Seção 4.3, a próxima etapa consistiu em realizar testes abrangendo todo o conjunto de dados e avaliando múltiplos modelos. Além disso, neste estágio, modelos treinados em língua portuguesa foram incorporados às análises.

Para todos os modelos, procedeu-se ao ranqueamento dos conjuntos de convênios, empregando os critérios de avaliação de acurácia para o top 1 e top 5, conforme abordado na Seção 2.6.1, além dos valores estatísticos descritos na Seção 2.6. Este processo foi aplicado ao conjunto de dados total composto por 1806 publicações.

As Tabelas 5.8 e 5.9 apresentam os resultados obtidos para a métrica de acurácia top-N, com colunas referentes ao resultado de top-1 e top-5. Vale ressaltar, que os modelos utilizados para a avaliação dos métodos de referência (Seção 5.3) também foram incorporados na etapa de modelagem do trabalho.

Tabela 5.8 – Resultados modelos em inglês.

Modelo	Top 1	Top 5
DistilRoBERTa (TRANSFORMERS, 2023a)	2,7%	10,4%
MiniLM L6 (TRANSFORMERS, 2023c)	12,2%	24,3%
MiniLM L12 (TRANSFORMERS, 2023b)	12,6%	20,7%
mpnet (TRANSFORMERS, 2023d)	7,7%	20,7%
DistilBERT (TRANSFORMERS, 2023e)	9,5%	18,5%
qa DistilBERT (TRANSFORMERS, 2023f)	9,0%	21,2%
qa MiniLM L6 (TRANSFORMERS, 2023g)	7,2%	25,7%
qa mpnet (TRANSFORMERS, 2023h)	4,1%	9,9%

Tabela 5.9 – Resultados modelos em português.

Modelo	Top 1	Top 5
BERTimbau-v1 (MELO; SANTOS; DIAS, 2023)	13,06%	25,23%
BERTimbau-v0 (MELO; SANTOS; DIAS, 2023)	13,06%	25,23%
BERTimbau (MELO, 2023b)	12,16%	21,17%
BERTimbau (MELO, 2023a)	14,41%	28,38%

Segundo a Tabela 5.8, os modelos em língua inglesa com maior índice de top-1, ou seja, que mais classificaram corretamente as publicações relacionadas ao seu convênio de origem foram os modelos MiniLM L6 (Transformers (2023c)) e MiniLM L12 (Transformers (2023b)) , com 12,2% e 12,6% de acerto. Considerando a porcentagem de acerto para o índice de top-5, atingiram os maiores índices os modelos MiniLM L6 (Transformers (2023c)) e qa MiniLM L6 (Transformers (2023g)), com 24,3% e 25,7%, respectivamente. Os resultados indicados pela cor vermelha são aqueles com menor índice de acerto no índice top-1 e no índice top-5

Para os resultados em português que são métodos derivados do modelo BERTimbau Souza, Rodrigo Nogueira e Lotufo (2020b) dessa forma contendo a mesma nomeclatura, mostrados pela Tabela 5.9, o modelo com o maior desempenho foi o modelo BERTimbau Melo (2023a), destacado os valores em azul, com 14,41% para a acurácia top-1 e 28,29% para a acurácia top-5 de acertos. Já os resultados indicados em vermelho são referentes ao modelo com o pior desempenho nessa etapa. O modelo de BERTimbau Melo (2023b) foi o que obteve o pior desempenho para o ranqueamento.

Os resultados dos modelos pré-treinados em língua portuguesa, conforme Tabela 5.9 mostram que em relação aos modelos pré-treinados em língua inglesa, ou em múltiplas línguas, o desempenho top-1 aumentou em todos os modelos e o desempenho top-5 aumentou em 3 dos 4 modelos mostrados. É importante ressaltar que o modelo BERTimbau (Melo (2023a)) obteve o maior índice para ambos os índices, o top-1 e top-5, com 14,41% de acertos no ranqueamento para a primeira posição e 28,38% para o ranqueamento de top-5.

Após o ranqueamento dos grupos de convênio com base na acurácia top-N, procedeu-se à avaliação de desempenho de todos os modelos, conforme mencionado na Seção 2.6. Isso envolveu a obtenção de métricas essenciais, como *precision*, *recall*, *F1-Score* e *accuracy* para todos os modelos. Os resultados para os modelos pré-treinados em língua inglesa ou multi-lingual podem ser observados na Tabela 5.10, enquanto os resultados dos modelos pré-treinados em português estão disponíveis na Tabela 5.11.

Tabela 5.10 – Métricas de avaliação modelos pré-treinados.

Modelo	EER (equal error rate)	Limiar EER	Precision	Recall	F1-score	Accuracy
DistilRoBERTa (Transformers (2023a))	0,41004	0,69972	0,00647	0,59002	0,01280	0,58991
MiniLM L6 (Transformers (2023c))	0,37780	0,62180	0,00740	0,62224	0,01462	0,62217
MiniLM L12 (Transformers (2023b))	0,39680	0,56035	0,00682	0,60265	0,01349	0,60311
mpnet (Transformers (2023d))	0,38928	0,70847	0,00704	0,61023	0,01392	0,61058
DistilBERT (Transformers (2023e))	0,38476	0,88766	0,00718	0,61466	0,01419	0,61520
qa DistilBERT (Transformers (2023f))	0,37467	0,66362	0,00749	0,62476	0,01480	0,62527
qa MiniLM L6 (Transformers (2023g))	0,38159	0,68519	0,00728	0,61845	0,01439	0,61838
qa mpnet (Transformers (2023h))	0,41377	0,74782	0,00637	0,58623	0,01260	0,58623

Tabela 5.11 – Métricas de avaliação nos modelos pré-treinados em língua portuguesa.

Modelo	EER (equal error rate)	Limiar EER	Precision	Recall	F1-score	Accuracy
BERTimbau-v1 (MELO; SANTOS; DIAS, 2023)	0,39926	0,682713	0,00676	0,60076	0,01337	0,60072
BERTimbau-v0 (MELO; SANTOS; DIAS, 2023)	0,39737	0,685090	0,00681	0,60202	0,01346	0,60260
BERTimbau (MELO, 2023b)	0,45484	0,686231	0,00539	0,54454	0,01067	0,54515
BERTimbau (MELO, 2023a)	0,39723	0,730647	0,00682	0,60265	0,01349	0,60289

As Tabelas 5.10 e 5.11 mostram as métricas citadas na Seção 2.6. Por meio dos valores destacados na Tabela 5.10, o modelo que melhor performou em todos os quesitos, foi um modelo pré-treinado em diversas linguagens. Já a Tabela 5.11, que demonstra os resultados em português, todos os modelos tiveram resultados muito próximo em todos as estatísticas de desempenho, exceto o modelo BERTimbau Melo (2023b), destacado em vermelho.

Considerando as Tabelas 5.8 e 5.9 e as Tabelas 5.10 e 5.11, percebe-se que na avaliação de desempenho de ranqueamento, principalmente na acurácia Top-N, a performance dos modelos pré-treinados em língua portuguesa tiveram uma maior eficiência. Ao considerar as demais métricas, percebe-se que o desempenho dos modelos em língua portuguesa não variaram dos modelos pré-treinados em inglês ou multi-lingual.

5.5 Treinamento

Com base na avaliação e estudo dos modelos de *Transformers* para a similaridade semântica, o próximo passo consistiu no treinamento do modelo escolhido com o conjunto de dados preparado, conforme detalhado na Seção 5.2. O treinamento é importante para tentar melhorar o desempenho do modelo utilizado-se de um conjunto de dados com res-

posta já definida, além de avaliar a capacidade do modelo de identificação de publicações correlacionadas na Seção 3 do DOU.

Para iniciar o processo do treinamento, a primeira etapa consiste na divisão do conjunto de dados, conforme descrito na Seção 5.2, em dados de treino e teste. Essa divisão é realizada aleatoriamente, com 70% dos dados separados para o treinamento e os 30% restantes separados para a realização dos testes posteriormente.

Após a separação dos dados de treinamento, utiliza-se o método de validação cruzada conhecido como *k-fold*, previamente abordado na Seção 4.5.2. Este método é aplicado para diferentes valores de k, buscando ter certeza que o treinamento é consistente independente do sorteio dos conjuntos de treinamento e testes. Dessa forma, observa-se se é necessário um nível de granularidade maior e com maior custo para o treinamento.

Ao realizar a separação das iterações é realizado o treinamento do método, utilizando uma função de perda. A função de perda permite o programa definir se o modelo está melhorando ou piorando ao longo do treinamento e ajustar-se conforme necessário. Nesse contexto, entre as funções mencionadas na Seção 4.5.3, a principal escolhida é a *CosineSimilarityLoss* para realizar o treinamento.

Durante o treinamento, são feitos testes com os valores do conjunto de testes de cada conjunto de treino. Após a conclusão do treinamento de cada Iteração da validação cruzada, é feito o teste final utilizando os dados de teste separados no início do processo de treino, obtendo-se os resultados finais. É importante notar que, em uma situação de validação cruzada, o valor de maior importância normalmente é a média dos resultados das iterações e não os valores individual de cada iteração.

O modelo principal a ser analisado foi o modelo BERTimbau (Melo (2023a)) . Conforme visto na Seção 5.4 é um método pre-treinado em português, com o melhor desempenho entre os modelos avaliados na métrica de acurácia Top-1 e Top-5, além de ser o modelo pré-treinado em português com a melhor performance nas demais métricas analisadas. O ranqueamento Top-N é um importante fator de decisão nesse momento porque define a acurácia global do modelo, determinando os conjuntos de publicações mais revelantes para cada publicação de convênio.

Com as diretrizes de treinamento estabelecidas anteriormente, procedeu-se ao treinamento para diferentes quantidades de iterações, visando avaliar o comportamento do sistema em relação à granularidade dos dados de treino. Os resultados dos testes realizados podem ser observados nas Tabelas 5.12, 5.13 e 5.14 a seguir:

Tabela 5.12 – Resultados do treinamento com 3 iterações.

Iteração	EER (equal error rate)	Limiar EER	Precision	Recall	F1-score	Accuracy
Iteração 1	0,52462	1,00000	0,00452	0,17722	0,00882	0,82099
Iteração 2	0,49609	1,00000	0,00424	0,18143	0,00828	0,80469
Iteração 3	0,44729	0,00727	0,00555	0,55274	0,01099	0,55269
Média	0,48933	0,66909	0,00477	0,30380	0,00937	0,72612
Modelo base	0,39843	0,73028	0,00677	0,60127	0,01340	0,60189

Tabela 5.13 – Resultados do treinamento com 5 iterações.

Iteração	EER (equal error rate)	Limiar EER	Precision	Recall	F1-score	Accuracy
Iteração 1	0,48411	1,00000	0,00917	0,00422	0,00578	0,99347
Iteração 2	0,50886	1,00000	0,01022	0,02743	0,01489	0,98369
Iteração 3	0,49063	1,00000	0,00362	0,00844	0,00507	0,98511
Iteração 4	0,51902	1,00000	0,00500	0,01266	0,00717	0,98424
Iteração 5	0,49927	1,00000	0,00442	0,14768	0,00858	0,84661
Média	0,50038	1,00000	0,00649	0,04008	0,00830	0,95862
Modelo base	0,39843	0,73028	0,00677	0,60127	0,01340	0,60189

Tabela 5.14 – Resultados do treinamento com 10 iterações.

Iteração	EER (equal error rate)	Limiar EER	Precision	Recall	F1-score	Accuracy
Iteração 1	0,49643	1,00000	0,00486	0,04852	0,00883	0,95105
Iteração 2	0,51035	1,00000	0,00349	0,01688	0,00578	0,97390
Iteração 3	0,53337	0,01698	0,00391	0,45992	0,00774	0,47014
Iteração 4	0,52292	1,00000	0,00407	0,31646	0,00804	0,64911
Iteração 5	0,42814	0,00100	0,00600	0,57173	0,01187	0,57200
Iteração 6	0,48851	1,00000	0,00398	0,12236	0,00772	0,85854
Iteração 7	0,48768	1,00000	0,00446	0,06540	0,00834	0,93009
Iteração 8	0,50389	1,00000	0,00427	0,17722	0,00834	0,81051
Iteração 9	0,49013	1,00000	0,00314	0,02954	0,00568	0,95353
Iteração 10	0,49192	1,00000	0,00601	0,02532	0,00971	0,97679
Média	0,49533	0,80180	0,00442	0,18333	0,00821	0,81457
Modelo base	0,39843	0,73028	0,00677	0,60127	0,01340	0,60189

Nesta seção, foi realizado o treinamento do modelo com diferentes configurações de iterações, explorando diversas granularidades dos dados de treino. Os resultados detalhados desses treinamentos serão analisados na próxima seção.

5.6 Análise de resultados

Com a metologia de trabalho proposta em 4 foi possível realizar o levantamento bibliográfico usado de referência para o presente o trabalho, alcançando os resultados dados na Seção 5.1.

A base de dados elaborada consiste de 1.806 publicações, com 222 conjunto de convênios, com um desbalanceamento no número de publicações em um conjunto de convênios, conforme demonstrado pelas Figuras 5.18a, 5.18b e 5.18c e pela a Tabela 5.5. A tabela demonstra que o máximo de publicações associados a um convênio é de 84 publicações e o mínimo é de 1 publicação. A média de publicações de um agrupamento é de aproximadamente 7,13.

Após a preparação do conjunto de dados, foi realizada em um subconjunto da base consistida por 559 publicações e 77 conjunto de convênios, o estudo preliminar dos métodos de referência, *Passage Ranking* e *Semantic Textual Similarity*. O principal objetivo da etapa descrita na Seção 4.3 era determinar a viabilidade do objetivo proposto e avaliar os métodos de referência de maneira a obter uma *baseline* para o trabalho. Apesar dos baixos valores obtidos para a *precision* nesse estágio, procedeu-se as etapas seguintes esperando ser possível visualizar a mudança na métrica.

A Seção 5.4 referencia os resultados obtidos por diferentes modelos em inglês e português. Em um primeiro momento foi realizada a determinação da acurácia top-1 e top-5 para todos os modelos. A utilização dessa métrica é importante principalmente para provar que, em média, o modelo não está classificando de maneira aleatória. Outra função relevante da métrica em classificação de documentos é a identificação do conjunto de publicações que sejam relevantes para o extrato de convênio comparado.

Nos modelos treinados em inglês ou em diversas línguas, conforme Tabela 5.8, o maior valor de acurácia top-1 é de 12,6%, ou seja, nos 222 conjuntos de convênios, o modelo MiniLM L12 (Transformers (2023b)) acertou o conjunto de publicações mais similar em termos semânticos 27 vezes. Por acurácia top-5 (qa MiniLM L6 (Transformers (2023g))), obteve apenas 7,2% de acerto na acurácia top-1, enquanto na top-5 obteve 25,7%, ou seja, classificou o conjunto de convenio entre os 5 primeiros conjuntos de documentos relevantes em 57 casos. Por meio da Tabela 5.8 e a Tabela 5.9, é possível realizar que os modelos em português desempenharam melhor considerado o método da acurácia top-n.

Considerando os resultados dos modelos em português, vide Tabela 5.9, os modelos não variaram muito em relação ao valor da acurácia top-1 e top-5, mas pode-se destacar que o modelo BERTimbau (Melo (2023a)) que obteve 14,1% de acertos para acurácia top-1 e 28,38% de acertos para acurácia top-5.

Conforme mencionado na Seção 5.4, também foram obtidas as estatísticas descritas anteriormente na Seção 2.6. A *precision* refere-se a estatística que mede dentre as classificações positivas que o modelo está considerando, quantas estão corretas. A Tabela 5.10 mostra que os valores para a métrica ficou entre 0,00636 e 0,00748. Os valores de *precision* são baixos porque uma das características de classificação do conjunto de dados é a baixa quantidade de verdadeiros positivos (TP), dessa forma, quando comparados com todas as predições verdadeiras (TP e FP) o valor é muito pequeno.

Por outro lado, o *recall*, demonstra dentre todas as situações que são esperadas positivas, quantas estão corretas, e conforme as Tabelas 5.10 e 5.11, se ajusta melhor ao conjunto de dados do trabalho, em que a média da estatística é de 0,6087 para os modelos em inglês e 0,5875 para os modelos em português.

A *F1-score* é a média harmônica entre a *precision* e *recall*, dessa forma, como os valores encontrados para as duas métricas estão desbalanceados, o valor de *F1-score* encontrado é baixo, ficando na casa de 1%.

Por fim, a *accuracy* indica a performance geral do modelo, ou seja, de todas as classificações, quantas o modelo classificou corretamente, dessa forma, o fato de ter poucos verdadeiros positivos não impacta nessa métrica, e os valores encontrados tem média de 58,78% para os modelos em português e de 60,88% para os modelos em português.

Conforme mencionado na Seção 5.5, o próximo passo consistiu no treinamento do modelo em português que obteve os melhores desempenhos no ranqueamento top-1 e top-5. Os treinamentos foram conduzidos com variações no número de iterações durante a validação cruzada, permitindo uma exploração abrangente das configurações do modelo. Esta abordagem possibilitou uma análise ampla do desempenho em diferentes cenários e verificar qual o melhor situação de treinamento.

Os resultados das Tabelas 5.12, 5.13 e 5.14 mostram os resultados de cada Iteração do treinamento e a média de todas as Iterações e por fim a comparação com o modelo base.

Os resultados das tabelas demonstram que a *accuracy* foi a métrica com o maior otimização no treinamento, com 72,61% para o treinamento com 3 iterações, 95,86% para o treinamento com 5 iterações e 81,46% para o treinamento com 10 iterações, em comparação com os 60,19% do modelo de base. A *precision* para o treinamento com 3 iterações é de 0,48%, 5 iterações é de 0,65% e de 0,44% para o de 10 iterações. O *recall* é dado por 30,38% no treinamento de 3 iterações, 4,01% para o treinamento de 5 iterações e de 18,83% no treinamento de 10 iterações. Por fim, a *F1-score* é de 0,94% no treinamento de 3 iterações, 0,83% no treinamento de de 5 iterações e 0,82% no treinamento de 10 iterações.

Por meio dos resultados apresentados, é possível concluir que de apesar do modelo possuir médias melhores do treinamento para o desempenho em *accuracy* em relação ao modelo base, os outros parâmetros analisados demonstram resultados em média piores que o modelo base, resultando em um *trade-off* entre as métricas analisadas.

Isso pode ser justificado pela natureza de classificação do conjunto de dados. Os dados analisados apresentam uma quantidade de classificação muito pequena de verdadeiros positivos (TP) por conjunto de convênios, ou seja, publicações que pertencem ao convênio, o que resulta no desbalanceamento das métricas calculadas. A EER representa o ponto em que as taxas de falsos positivos e falsos negativos são iguais, ou seja é um limiar de decisão do sistema em relação a classificação do modelo. No treinamento, o limiar EER de decisão é

dado por um valor alto que acaba por classificar como não pertencente ao convênio a grande maioria das publicações.

É importante ressaltar que apesar dos valores do limiar EER são aproximadamente iguais a 1, e não exatamente 1, ou seja, a diferença obtida mostra que a decisão de não classificar uma publicação como negativa, o modelo não está simplesmente classificando como negativos todas as publicações analisadas.

Como exemplo do *trade-off*, pode-se citar a iteração 1 da tabela 5.13. Essa iteração possui 99,34% de *accuracy*, o maior valor do treinamento realizado. Por outro lado, tem o pior valor para o *recall*. Esse desempenho aconteceu porque nessa iteração classificou como não sendo pertencente do convênio quase todos os valores de entrada, conforme o limiar do EER de aproximadamente 1. Já a segunda iteração, apesar de ter o valor de *accuracy* de 98,36%, menor que a iteração 1, possui o valor de *precision* e *F1-score* maior que o modelo base, ou seja, nessa iteração, houve uma redução dos erros de classificação dos valores que são considerados positivos (FP).

A questão dos verdadeiros positivos é intrínseca ao problema proposto, pois há uma quantidade de documentos que não se correlacionam muito maior do que documentos correlacionados no DOU. Outro fator determinante no estudo são os *Transformers* utiliza a posição das palavras no textos. Ao utilizar publicações da seção 3 do DOU, algumas estruturas são comuns a vários documentos, como "Extrato de Convênio", "Prazo de Vigência", entre outros elementos, o que leva a uma falsa proximidade semântica para os modelos. Dessa forma, ao treinar o modelo da maneira proposta, é realizada uma classificação binária por meio do limiar EER no sentido de considerar extremamente próximo, ou extremamente diferentes, conforme a Figura 5.21.

DIÁRIO OFICIAL DA UNIÃO

Publicado em: 03/09/2019 | Edição: 170 | Seção: 3 | Página: 14

Órgão: Ministério da Defesa/Secretaria-Geral

EXTRATO DE CONVÊNIO

Espécie: Convênio Nº 882985/2019, Nº Processo: 60414000262201952, Concedente: MINISTERIO DA DEFESA, Convenente: MUNICIPIO DE CEREJEIRAS CNPJ nº 04914925000107, Objeto: Pavimentação Asfáltica em Via Urbana com Drenagem e Calçadas no Município de Cerejeiras/RO., Valor Total: R$ 1.245.988,00, Valor de Contrapartida: R$ 32.000,00, Valor a ser transferido ou descentralizado por exercício: 2019 - R$ 1.213.988,00, Crédito Orçamentário: Num Empenho: 2019NE800206, Valor: R$ 1.213.988,00, PTRES: 150099, Fonte Recurso: 0188000000, ND: 44425141, Vigência: 02/09/2019 a 12/08/2023, Data de Assinatura: 02/09/2019, Signatários: Concedente: ROBERTO DE MEDEIROS DANTAS CPF nº 483.922.198-72, Convenente: LISETE MARTH CPF nº 526.178.310-00.

(a) Extrato de convênio.

DIÁRIO OFICIAL DA UNIÃO

Publicado em: 24/12/2021 | Edição: 242 | Seção: 3 | Página: 3

Órgão: Ministério da Agricultura, Pecuária e Abastecimento/Gabinete da Ministra

EXTRATO DE TERMO ADITIVO

Espécie: Termo Aditivo de Alteração da Vigência Nº 000001/2021 ao Convênio Nº 889074/2019. Convenentes: Concedente: MINISTÉRIO DA AGRICULTURA, PECUARIA E ABASTECIMENTO, Unidade Gestora: 130005. Convenente: MUNICIPIO DE PAROBE, CNPJ nº 88372883000101. O Objeto é a Alteração de vigência do Convênio 889074/2019 para mais 180 dias, ou seja, 30 de junho de 2022. Valor Total: R$ 177.500,00, Valor de Contrapartida: R$ 17.500,00, Vigência: 22/12/2021 a 30/06/2022. Data de Assinatura: 31/12/2019. Signatários: Concedente: EDIMILSON ALVES, CPF nº 60608900168, Convenente: DIEGO DAL PIVA DA LUZ, CPF nº 007.648.140-95.

(b) Extrato de Termo Aditivo.

Figura 5.21 – Publicações de convênios distintas.

Fonte: (IMPRENSA NACIONAL, 2023)

A Figura 5.21 mostra uma publicação de Extrato de Convênio (Figura 5.21a) do convênio número 882985/2019, e a publicação de um Extrato de Termo Aditivo (Figura 5.21b) do convênio número 889074/2019. Ambas as publicações contêm entidades como "Convenentes", "Concedente", "Ministério", "Valor total", "Vigência", "Data de Assinatura", o que leva ao modelo classificar publicações que não são correlacionadas como pertencente a um conjunto de convênios.

6 Conclusões

O presente trabalho teve como objetivo principal realizar o estudo, análise e treinamento de modelos de *Transformers* previamente treinados para encontrar semelhança semântica entre os elementos de publicações contratuais da Seção 3 do Diário Oficial da União, estabelecendo um vínculo entre elas. Para atingir esse objetivo, foi desenvolvido um conjunto de dados composto por publicações com uma hierarquia de relacionamento entre elas e uma análise e avaliação dos diferentes modelos de *Transformers* disponíveis na literatura.

Na primeira etapa, foi realizado o levantamento bibliográfico de referência utilizado no presente trabalho, conforme mencionado na Seção 5.1.

A segunda etapa consistiu no estabelecimento do conjunto de dados elaborado com 1.806 publicações, com 222 conjuntos de convênios e conforme analisado na Seção 5.6, com uma grande diferença no número de publicações em cada conjunto de convênios e no tamanho de cada publicação. A característica de desbalanceamento entre a quantidade de elementos textuais da base que pertencem a um determinado convênio e a quantidade dos mesmos que não pertencem ao dado convenio é importante para explicar os resultados obtidos a seguir.

Na terceira etapa, conforme a Seção 4.3, realizou-se um estudo preliminar dos métodos de referência, em um subconjunto da base de dados, de maneira a determinar que mesmo com resultados desanimadores para algumas métricas ainda há viabilidade para continuar o trabalho estabelecido.

A partir da quarta etapa do trabalho, é apresentado diferentes modelos de *Transformers* pré-treinados em língua inglesa e língua portuguesa de maneira a avaliar o relacionamento semântico e verificar a possibilidade de estabelecer o vínculos entre publicações de convênios, licitações e contratos.

Na avaliação de acurácia top-n, em que é avaliado o desempenho do modelo no ranqueamento do conjunto de convênios mais próximo a publicação de busca, os modelos em português performaram melhor, atingindo o valor de 14,41% para a acurácia top 1 e de 20,38% para a acurácia top 5.

Por outro lado, os modelos em português não apresentaram desempenho superior aos modelos pré-treinados em inglês quando analisado as métricas de *precision*, *Accuracy*, *F1-score* e *recall*.

Dada a grande quantidade de predições verdadeiras comparadas com a quantidade de verdadeiros positivos, ou seja, publicações que realmente fazem parte daquele grupo de

convênio, os valores de *precision* obtidos são menores que 1%. Já os valores de *recall* não são afetados com tanta intensidade por esse desbalanceamento, pois é a identificação de todos os exemplos positivos que foram corretamente classificados em relação a todos os exemplos que são verdadeiros no conjunto de dados. A métrica teve média de 59% para os modelos pré-treinados em português e 61% para os modelos pré-treinados em inglês. A *F1-score*, conforme mencionado na Seção 5.6, é a média harmônica entre as métricas referidas, e por isso ficou em torno de 1%. A *accuracy*, por outro lado, avaliando os modelos pre-treinado, atingiu valores em média de 58,8% para os modelos pré-treinados em português e 60,9% para os modelos pré-treinados em inglês.

Concluída a etapa de avaliação dos modelos, selecionou-se o modelo a ser treinado com base no desempenho superior na métrica de *accuracy* top-n durante a análise. O treinamento foi realizado com a separação dos dados em 70% dos dados foram para treinar o modelo e 30% dos dados foram utilizados para os testes. O treinamento foi feito com o método de avaliação cruzada, e com diferentes níveis de granularidade.

Dessa forma, foram realizados treinamentos com o método de validação cruzada *K-fold*. Concluiu-se que a *accuracy* do modelo após o treinamento foi a métrica que melhorou em comparação ao modelo base. Isso se dá pelo o alto valor de limiar EER, que quando mais alto, menor o nível de classificação positiva, e menor a quantidade dos falsos positivos, aumentando então a *accuracy* desempenhada. Por outro lado, as demais métricas mostram que o modelo não melhorou seu desempenho nelas. O treinamento com o melhor comportamento de *accuracy* foi o com 5 iterações, com 95,86% de *accuracy*.

Para trabalhos futuros, é importante avaliar o desequilíbrio entre as publicações, visando aprimorar o aprendizado do modelo. Outras estratégias viáveis incluem explorar o treinamento de modelos adicionais, incorporando diferentes variáveis de treinamento. Uma abordagem interessante seria diversificar o conjunto de dados de treinamento, explorando diferentes níveis de gradação, a fim de evitar uma classificação binária e mitigar potenciais perdas nas métricas de avaliação.

É fundamental destacar que este trabalho busca aprimorar a pesquisa relacionada à aplicação de modelos de aprendizagem profunda em instrumentos contratuais públicos. Ele representa um ponto de partida em uma área ampla que ainda oferece uma vasta gama de oportunidades para exploração, como a determinação da hierarquia entre as publicações.

Referências

BRASIL. **Base de Dados de Publicações do DOU**. Acesso em: 14 ago. 2023. 2023a. Disponível em: <https://www.in.gov.br/acesso-a-informacao/dados-abertos/base-de-dados>. Citado na p. 16.

BRASIL. **Constituição da República Federativa do Brasil de 1988.** Acesso em: 29 nov. 2023. 1988. Disponível em: <https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm>. Citado na p. 13.

BRASIL. **DECRETO Nº 9.215, DE 29 DE NOVEMBRO DE 2017**. Acesso em: 26 de nov. 2023. 2017a. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2017/decreto/d9215.htm>. Citado na p. 14.

BRASIL. **Diário Oficial da União registra história do Brasil desde o Império**. Acesso em: 15 ago. 2023. Dez. 2017b. Disponível em: <https://www.gov.br/secretariageral/pt-br/noticias/2017/dezembro/diario-oficial-da-uniao-registra-historia-do-brasil-desde-o-imperio>. Citado na p. 16.

BRASIL. LEI Nº 14.133, DE 1º DE ABRIL DE 2021. **Diário Oficial da República Federativa do Brasil**, Brasília, DF, 2021. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14133.htm>. Citado nas pp. 16, 18, 19.

BRASIL. **Termo aditivo a convênio de adesão**. Acesso em: 25 jun.2023. 2023b. Disponível em: <https://www.gov.br/previc/pt-br/licenciamento-e-habilitacao/entidades-planos-e-patrocinadores/termo-aditivo-a-convenio-de-adesao>. Citado na p. 20.

BROWNE, M. W. Cross-Validation Methods. **Journal of Mathematical Psychology**, v. 44, n. 1, p. 108–132, 2000. ISSN 0022-2496. DOI: https://doi.org/10.1006/jmps.1999.1279. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0022249699912798>. Citado na p. 44.

CANÇADO, M. **Manual de semântica: noções básicas e exercícios**. 2. ed. Belo Horizonte: Editora UFMG, 2008. ISBN 9788570416803. Citado nas pp. 14, 15.

CAPELA, M. V.; CAPELA, J. M. Elaboração de gráficos box-plot em planilhas de cálculo. In: CONGRESSO de matemática aplicada e computacional da região sudeste–cnmac Sudeste. 2011. v. 1. Citado na p. 49.

CARVALHO, L. R. D.; LOPES, F.; CHAVES, J.; LIMA, M.; DEUS, F. G. D.; PAUNGARTHEM, A. von; VIDAL, F. Deep-vacuity: A Proposal of a Machine Learning Platform based on High-performance Computing Architecture for Insights on Government of Brazil Official Gazettes. In: INSTICC. PROCEEDINGS of the 18th International Conference

on Web Information Systems and Technologies - Volume 1: WEBIST, SciTePress, 2022. P. 136–143. ISBN 978-989-758-613-2. DOI: 10.5220/0011532500003318. Citado na p. 35.

CER, D.; DIAB, M.; AGIRRE, E.; LOPEZ-GAZPIO, I.; SPECIA, L. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In: PROCEEDINGS of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Vancouver, Canada: Association for Computational Linguistics, ago. 2017. DOI: 10.18653/v1/S17-2001. Disponível em: <https://aclanthology.org/S17-2001>. Citado na p. 42.

CLARKE, N.; FURNELL, S.; REYNOLDS, P. Biometric Authentication for Mobile Devices. **Proceeding of the 3rd Australian Information Warfare and Security Conference**, jan. 2002. Citado na p. 34.

CONNECTED PAPERS. **Connected Papers**. Acesso em: 10 out. 2023. 2023. Disponível em: <https://www.connectedpapers.com/main/93d63ec754f29fa22572615320afe0521f7ec66d/Sentence%5C%20BERT%5C%3A-Sentence-Embeddings-using-Siamese-BERT%5C%20Networks/graph>. Citado na p. 40.

DEEP VACUITY. 2021. Disponível em: <https://deepvacuity.cic.unb.br/#/login>. Citado na p. 17.

DEPARTAMENTO DE LOGÍSTICA E SERVIÇOS GERAIS. Ajuste do Plano de Trabalho, Termo Aditivo e Prorroga Ofício. **Portal dos Convênios - SICONV**, 2013. Disponível em: <https://www.gov.br/plataformamaisbrasil/pt-br/manuais-e-cartilhas/arquivos-e-imagens/convenente_concedente_ajuste_plano_trabalho_termo_aditivo_prorroga_oficio.pdf>. Citado na p. 21.

DEZA, M. M.; DEZA, E. **Encyclopedia of Distances**. Germany: Springer Berlin Heidelberg, 2009. Citado na p. 24.

DIRETORIA DE COMPRAS, CONTRATOS E CONVÊNIOS. **Manual de contratos e convênios**. Acesso em: 15 ago. 2023. Universidade Federal do Sul e Sudeste do Pará - UNIFESSPA. 2018. Disponível em: <https://proad.unifesspa.edu.br/images/subunidades/dco/dicc/manuais/manual-v3.pdf>. Citado na p. 20.

FONSECA, E. R.; BORGES DOS SANTOS, L.; CRISCUOLO, M.; ALUÍSIO, S. M. Visão Geral da Avaliação de Similaridade Semântica e Inferência Textual. **Linguamática**, v. 8, n. 2, p. 3–13, dez. 2016. Disponível em: <https://linguamatica.com/index.php/linguamatica/article/view/v8n2-1>. Citado na p. 37.

FORTINI, C.; MOTTA, F. Corrupção nas licitações e contratações públicas: sinais de alerta segundo a Transparência Internacional. **A&C : Revista de Direito Administrativo & Constitucional**, 2003. ISSN 1516-3210. DOI: http://dx.doi.org/10.21056/aec.v16i64.240. Citado na p. 13.

HOFSTÄTTER, S.; LIN, S.-C.; YANG, J.-H.; LIN, J.; HANBURY, A. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. In: PROC. of SIGIR. 2021. Citado na p. 51.

IBM. **What is natural language processing?** Acesso em: 17 ago. 2023. 2023. Disponível em: <https://www.ibm.com/topics/natural-language-processing>. Citado na p. 22.

IMPRENSA NACIONAL. **Diário Oficial da União**. Acesso em: 6 nov. 2023. 2023. Disponível em: <https://www.in.gov.br/leiturajornal>. Citado nas pp. 18–21, 42, 60.

JONES, K. S. Natural language processing: a historical review. **University of Cambridgee**, out. 2001. Acesso em: 24 abr. 2023. Citado na p. 22.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing**. 3. ed. Califórnia, 2023. Acesso em: 18 ago. 2023. Citado nas pp. 25–27, 29–31.

LEE, J.; LEE, D.; LEE, Y.-C.; HWANG, W.-S.; KIM, S.-W. Improving the accuracy of top-N recommendation using a preference model. **Information Sciences**, v. 348, p. 290–304, 2016. ISSN 0020-0255. DOI: https://doi.org/10.1016/j.ins.2016.02.005. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0020025516300524>. Citado na p. 33.

LIMA, M.; SILVA, R.; LOPES DE SOUZA MENDES, F.; R. DE CARVALHO, L.; ARAUJO, A.; BARROS VIDAL, F. de. Inferring about fraudulent collusion risk on Brazilian public works contracts in official texts using a Bi-LSTM approach. In: FINDINGS of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, nov. 2020. P. 1580–1588. DOI: 10.18653/v1/2020.findings-emnlp.143. Disponível em: <https://aclanthology.org/2020.findings-emnlp.143>. Citado nas pp. 14, 35.

MACÊDO, S. **A importância dos Diários Oficiais**. Acesso em: 15 ago. 2023. Nov. 2018. Disponível em: <https://al.se.leg.br/a-importancia-dos-diarios-oficiais/>. Citado na p. 16.

MAJUMDER, G.; PAKRAY, P.; GELBUKH, A.; PINTO, D. Semantic Textual Similarity Methods, Tools, and Applications: A Survey. **Computación y Sistemas**, v. 20, n. 4, p. 647–665, dez. 2016. Disponível em: <https://doi.org/10.13053/cys-20-4-2506>. Citado nas pp. 23, 37.

MATHWORKS. **Long Short-term Memory (LSTM)**. Acesso em: 14 dez. 2023. 2023. Disponível em: <https://www.mathworks.com/discovery/lstm.html#:~:text=LSTMs%20are%20predominantly%20used%20to,speech%20recognition%2C%20and%20video%20analysis.>. Citado na p. 35.

MELO, R. **bert-large-portuguese-cased-sts**. Acesso em:27/11/2023. 2023a. Disponível em: <https://huggingface.co/rufimelo/bert-large-portuguese-cased-sts>. Citado nas pp. 53–55, 57.

MELO, R. **Legal-BERTimbau-sts-large**. Acesso em:27/11/2023. 2023b. Disponível em: <https://huggingface.co/rufimelo/Legal-BERTimbau-sts-large>. Citado nas pp. 53, 54.

MELO, R.; SANTOS, P. P. A.; DIAS, P. J. A Semantic Search System for Supremo Tribunal de Justiç. In. Citado nas pp. 53, 54.

MICROSOFT. **Modelo de validação cruzada**. Acesso em: 8 out. 2023. 2023. Disponível em: <https://learn.microsoft.com/pt-br/azure/machine-learning/component-reference/cross-validate-model?view=azureml-api-2>. Citado na p. 44.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient Estimation of Word Representations in Vector Space. **CoRR**, abs/1301.3781, 2013. Disponível em: <http://dblp.uni-trier.de/db/journals/corr/corr1301.html#abs-1301-3781>. Citado na p. 23.

MORAES PINTO, D. C. de; COELHO, F. A. C.; LIMA CABRAL, M. P. de; RIBEIRO, R. M. P. **Introdução à semântica**. Rio de Janeiro: Fundação Cecierj, 2016. ISBN 9788545800316. Citado na p. 14.

NOGUEIRA, R. F.; JIANG, Z.; LIN, J. Document Ranking with a Pretrained Sequence-to-Sequence Model. **CoRR**, abs/2003.06713, 2020. Disponível em: <https://arxiv.org/abs/2003.06713>. Citado na p. 42.

OLIVEIRA, N. **Nova Lei de Licitações é esperança contra corrupção e desperdício de verbas**. Edição: Agência Senado. Acesso em: 27 out. 2023. Abr. 2021. Disponível em: <https://www12.senado.leg.br/noticias/infomaterias/2020/12/nova-lei-de-licitacoes-e-esperanca-contra-corrupcao-e-desperdicio-de-verbas>. Citado na p. 13.

PENNINGTON, J.; SOCHER, R.; MANNING, C. GloVe: Global Vectors for Word Representation. In: PROCEEDINGS of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, out. 2014. Disponível em: <https://aclanthology.org/D14-1162>. Citado nas pp. 23, 24.

PETERS, M. E.; NEUMANN, M.; IYYER, M.; GARDNER, M.; CLARK, C.; LEE, K.; ZETTLEMOYER, L. Deep contextualized word representations. **CoRR**, abs/1802.05365, 2018. Disponível em: <http://arxiv.org/abs/1802.05365>. Citado nas pp. 23, 24.

PORTAL DA TRANSPARÊNCIA. **Convênios e outros acordos**. Acesso em: 28 mai. 2023. 2023a. Disponível em: <https://portaldatransparencia.gov.br/entenda-a-gestao-publica/convenios-e-outros-acordos>. Citado na p. 17.

PORTAL DA TRANSPARÊNCIA. **Licitações e contratações**. Acesso em: 28 mai. 2023. 2023b. Disponível em: <https://portaldatransparencia.gov.br/entenda-a-gestao-publica/licitacoes-e-contratacoes>. Citado na p. 18.

POWERS, D. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. **Mach. Learn. Technol.**, v. 2, jan. 2008. Citado na p. 32.

RAMALHO, H. V. A interface semântico-sintática na construção dos textos escolares. **Cadernos do CNLF**, v. 12, n. 9, 2008. Citado na p. 15.

REIMERS, N.; GUREVYCH, I. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, nov. 2020. Disponível em: <https://arxiv.org/abs/2004.09813>. Citado na p. 36.

REIMERS, N.; GUREVYCH, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. **CoRR**, abs/1908.10084, 2019. arXiv: 1908.10084. Disponível em: <http://arxiv.org/abs/1908.10084>. Citado nas pp. 35, 45, 46.

RENTE NETO, F. A corrupção no Brasil contemporâneo: um estudo sobre os crimes políticos federais no Brasil (2004-2015). **Em Tempo de Histórias**, v. 1, n. 39, dez. 2021. DOI: 10.26512/emtempos.v1i39.39903. Disponível em: <https://periodicos.unb.br/index.php/emtempos/article/view/39903>. Citado na p. 13.

RESNIK, P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In: INTERNATIONAL Joint Conference on Artificial Intelligence. 1995. Disponível em: <https://api.semanticscholar.org/CorpusID:1752785>. Citado na p. 36.

SANH, V.; DEBUT, L.; CHAUMOND, J.; WOLF, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. **CoRR**, abs/1910.01108, 2019. arXiv: 1910.01108. Disponível em: <https://dblp.org/rec/journals/corr/abs-1910-01108.bib>. Citado na p. 35.

SAS. **Natural Language Processing (NLP)**. Acesso em: 14 ago. 2023. 2023. Disponível em: <https://www.sas.com/en_nz/insights/analytics/what-is-natural-language-processing-nlp.html>. Citado na p. 22.

SCIKIT LEARN. **3.1. Cross-validation: evaluating estimator performance**. Acesso em: 8 out. 2023. 2023. Disponível em: <https://scikit-learn.org/stable/modules/cross_validation.html>. Citado na p. 45.

SLIMANI, T. Description and Evaluation of Semantic Similarity Measures Approaches. **International Journal of Computer Applications**, Foundation of Computer Science, v. 80, n. 10, p. 25–33, out. 2013. DOI: 10.5120/13897-1851. Disponível em: <https://doi.org/10.5120%2F13897-1851>. Citado nas pp. 23, 36.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In: CERRI, R.; PRATI, R. C. (Ed.). **Intelligent Systems**. Cham: Springer International Publishing, 2020a. P. 403–417. ISBN 978-3-030-61377-8. Citado na p. 36.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: 9TH Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear). 2020b. Citado na p. 53.

STF. **A publicidade dos atos e decisões administrativos**. 2023. Citado na p. 13.

TEIXEIRA, A. B.; REHBEN-SATLHER, A. G.; RODRIGUES, M. R. Percepções sociais sobre a corrupção política no Brasil: práticas corruptas versus atuação dos órgãos de controle. **Colombia Internacional [En línea]**, n. 105, 2021. Citado na p. 13.

TRANSFORMERS, S. **all-distilroberta-v1**. Acesso em:27/11/2023. 2023a. Disponível em: <https://huggingface.co/sentence-transformers/all-distilroberta-v1>. Citado nas pp. 36, 53, 54.

TRANSFORMERS, S. **all-MiniLM-L12-v2**. Acesso em:27/11/2023. 2023b. Disponível em: <https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2>. Citado nas pp. 36, 53, 54, 57.

TRANSFORMERS, S. **all-MiniLM-L6-v2**. Acesso em:27/11/2023. 2023c. Disponível em: <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>. Citado nas pp. 36, 53, 54.

TRANSFORMERS, S. **all-mpnet-base-v2**. Acesso em:27/11/2023. 2023d. Disponível em: <https://huggingface.co/sentence-transformers/all-mpnet-base-v2>. Citado nas pp. 36, 53, 54.

TRANSFORMERS, S. **msmarco-distilbert-base-tas-b**. Acesso em:27/11/2023. 2023e. Disponível em: <https://huggingface.co/sentence-transformers/msmarco-distilbert-base-tas-b>. Citado nas pp. 53, 54.

TRANSFORMERS, S. **multi-qa-distilbert-cos-v1**. Acesso em:27/11/2023. 2023f. Disponível em: <https://huggingface.co/sentence-transformers/multi-qa-distilbert-cos-v1>. Citado nas pp. 36, 53, 54.

TRANSFORMERS, S. **multi-qa-MiniLM-L6-cos-v1**. Acesso em:27/11/2023. 2023g. Disponível em: <https://huggingface.co/sentence-transformers/multi-qa-MiniLM-L6-cos-v1>. Citado nas pp. 36, 53, 54, 57.

TRANSFORMERS, S. **multi-qa-mpnet-base-dot-v1**. Acesso em:27/11/2023. 2023h. Disponível em: <https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1>. Citado nas pp. 36, 53, 54.

TRIBUNAL DE CONTAS DA UNIÃO. Licitações e contratos : orientações e jurisprudência do TCU. **Brasília : TCU, Secretaria-Geral da Presidência : Senado Federal, Secretaria Especial de Editoração e Publicações**, 2010. Acesso em: 15 ago. 2023. Citado na p. 19.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. Attention Is All You Need. **CoRR**, abs/1706.03762, 2017. Disponível em: <http://arxiv.org/abs/1706.03762>. Citado nas pp. 23, 24.

WANG, J.; DONG, Y. Measurement of Text Similarity: A Survey. **Information**, v. 11, n. 9, 2020. ISSN 2078-2489. Disponível em: <https://www.mdpi.com/2078-2489/11/9/421>. Citado na p. 23.

WANG, W.; WEI, F.; DONG, L.; BAO, H.; YANG, N.; ZHOU, M. **MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers**. 2020a. arXiv: 2002.10957 [cs.CL]. Citado na p. 51.

WANG, W.; WEI, F.; DONG, L.; BAO, H.; YANG, N.; ZHOU, M. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. **CoRR**, abs/2002.10957, 2020b. arXiv: 2002.10957. Disponível em: <https://dblp.org/rec/journals/corr/abs-2002-10957.bib>. Citado na p. 35.

WIKIPEDIA CONTRIBUTORS. **F-score — Wikipedia, The Free Encyclopedia**. 2023. Online; Acesso em: 19 jul. 2023. Disponível em: <https://en.wikipedia.org/w/index.php?title=F-score&oldid=1148225663>. Citado na p. 33.