UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Agrupamento Semântico no Contexto de Emendas Legislativas

Pedro Lucas Castro de Andrade

São Carlos – SP

Agrupamento Semântico no Contexto de Emendas Legislativas

Pedro Lucas Castro de Andrade

Orientador: **Prof. Dr. André Carlos Ponce de Leon Ferreira de Carvalho**

Coorientadoras: **Profas. Dras. Ellen Souza e Nádia Silva**

Monografia final de conclusão de curso do Departamento de Sistemas de Computação do Instituto de Ciências Matemáticas e de Computação – ICMC-USP, para obtenção do título de Engenheiro de Computação.
Área de Concentração: Inteligência Artificial, Processamento de Linguagem Natural

USP – São Carlos
Novembro de 2024

Andrade, Pedro Lucas Castro de
Agrupamento Semântico no Contexto de Emendas Legislativas / Pedro Lucas Castro de Andrade. - São Carlos - SP, 2024.
42 p.; 29,7 cm.

Orientador: André Carlos Ponce de Leon Ferreira de Carvalho.
Coorientadoras: Ellen Souza e Nádia Silva.
Monografia (Graduação) - Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos - SP, 2024.

1. Inteligência Artificial. 2. Processamento de Linguagem Natural. 3. Agrupamento Semântico. 4. Texto Legislativo. 5. *Large Language Model*. I. Carvalho, André Carlos Ponce de Leon Ferreira de. II. Instituto de Ciências Matemáticas e de Computação (ICMC/USP). III. Agrupamento Semântico no Contexto de Emendas Legislativas.

Aos meus pais, que se desdobraram imensamente para que essa conquista fosse possível.

AGRADECIMENTOS

Por primeiro, agradeço aos meus pais, José Marcelo e Dulciene, por ensinarem-me os fundamentos da vida, em especial, a dedicação, a resiliência, o amor e a fé, e por proporcionarem-me, sem medirem esforços, anos em que pude me dedicar aos estudos e à graduação. À minha irmã, Vitória, quem presenteia-me com sua alegria e leveza, e a todos os meus familiares que estiveram presentes ao longo destes anos.

Ao Prof. Dr. André de Carvalho e às Profas. Dras. Ellen Souza e Nádia Silva pela orientação e coorientação neste trabalho e pela oportunidade de trabalhar no projeto Ulysses, onde obtive contato com a pesquisa científica e, além disso, com colegas de graduação e pós-graduação, com os quais recebi valiosos ensinamentos. À Universidade de São Paulo, a qual trouxe-me a São Carlos, permitindo que vivesse anos de estudos, convivência com novas pessoas, amadurecimento pessoal e profissional.

Aos amigos que fiz nesse tempo, sobretudo Alexandre, Fabrício, Gabriel, Lucas, Luigi, Michel, Rafael e Vinícius, as conversas durante os almoços, as trocas de metas e objetivos, as horas intensas de estudos em vésperas de provas e todos os outros momentos de convivência fizeram com que esse período passasse de maneira mais amena.

Por fim, expresso o mais importante dos agradecimentos, a Deus, que é Pai - a fonte de toda vida e Amor por princípio -, Filho - a Verdade na qual me fortaleço - e Espírito Santo - o Paráclito que auxilia-me nos momentos de provação.

"O trabalho é a vocação inicial do homem, é uma bênção de Deus, e enganam-se lamentavelmente os que o consideram um castigo." (São Josemaria Escrivá)

RESUMO

ANDRADE, P. L. C. DE. **Agrupamento Semântico no Contexto de Emendas Legislativas**. 2024. 42 f. Monografia (Graduação) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

Este trabalho explora a realização de agrupamento semântico de emendas legislativas brasileiras por meio de *Large Language Models*. Considerando o número elevado de emendas redigidas anualmente na Câmara dos Deputados Federal e no Senado Federal e a consequente carga horária massiva despendida pelas equipes responsáveis por agrupar tais documentos, o uso de técnicas automatizadas para análise e organização eficientes apresenta-se como uma alternativa benéfica às casas legislativas. A partir dos modelos GPT-4o e GPT-4o-mini da OpenAI, realizaram-se experimentos visando agrupar emendas conforme sua similaridade semântica. O estudo comparou abordagens com e sem pré-processamento textual, variando o valor da temperatura do modelo para avaliar o impacto deste parâmetro sobre a qualidade do resultado final. Como forma de validar a eficácia dos grupos formados, as métricas de precisão, *recall* e F1 foram aplicadas. Constatou-se que o modelo GPT-4o apresentou desempenho superior, sobretudo com temperaturas inferiores. Não obstante, este estudo foi limitado pelos custos de infraestrutura e a propensão dos modelos a alucinarem durante sua execução. Os resultados obtidos indicam potencial dos modelos em auxiliar a tarefa de agrupar emendas, oferecendo subsídios para futuras melhorias no uso de Inteligência Artificial no contexto legislativo brasileiro.

Palavras-chave: Inteligência Artificial, Processamento de Linguagem Natural, Agrupamento Semântico, Texto Legislativo, *Large Language Model*.

ABSTRACT

ANDRADE, P. L. C. DE. **Agrupamento Semântico no Contexto de Emendas Legislativas**. 2024. 42 f. Monografia (Graduação) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

This study explores the semantic clustering of Brazilian legislative amendments through Large Language Models. Given the high number of amendments drafted annually in the Brazilian Federal Chamber of Deputies and the Federal Senate, and the consequent extensive hours spent by teams responsible for grouping these documents, the use of automated techniques for efficient analysis and organization presents itself as a beneficial alternative for legislative bodies. Using OpenAI's GPT-4o and GPT-4o-mini models, experiments were conducted to group amendments according to their semantic similarity. The study compared approaches with and without text preprocessing, varying the model's temperature parameter to assess its impact on result quality. To validate the effectiveness of the formed clusters, precision, recall, and F1 metrics were applied. It was found that the GPT-4o model achieved superior performance, especially with lower temperatures. Nonetheless, this study was constrained by infrastructure costs and the models' tendency to hallucinate during execution. The results indicate the models' potential to assist in clustering amendments, providing insights for future improvements in the application of Artificial Intelligence within the Brazilian legislative context.

Key-words: Artificial Intelligence, Natural Language Processing, Semantic Clustering, Legislative Text, Large Language Model.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

API ..... *Application Programming Interface*
BERT .... *Biredctional Encoder Representations from Transformers*
BM25 .... *Best Match 25*
CNN ..... Rede Neural Convolucional
GPU ..... *Graphic Processing Unit*
IA ..... Inteligência Artificial
ICMC .... Instituto de Ciências Matemáticas e de Computação
LLM ..... *Large Language Model*
NER ..... Entidades Nomeadas
NLTK .... *Natural Language Toolkit*
PEC ..... Proposta de Emenda à Constituição
PL ..... Projeto de Lei
PLN ..... Processamento de Linguagem Natural
RLHF .... *Reinforcement Learning from Human Feedback*
RLSP .... Removedor de Sufixos da Língua Portuguesa
RNN ..... Rede Neural Recorrente
SBERT ... *Sentence Transformer*
USP ..... Universidade de São Paulo

SUMÁRIO

Capítulo 1

INTRODUÇÃO

1.1 Motivação e Contextualização

Dentre as atividades desempenhadas no Poder Legislativo Brasileiro, há o longo processo de elaboração, análise e votação de Projeto de Lei (PL) e de Proposta de Emenda à Constituição (PEC), documentos importantes para a atualização e aprimoramento do sistema legal do país. Durante esse trâmite, são propostos pelos parlamentares documentos chamados de emendas, cuja finalidade é modificar, adicionar ou remover disposições do texto original do PL ou da PEC. Uma vez criadas, as emendas necessitam de avaliação para averiguar sua admissibilidade e, em seguida, são debatidas e votadas por parlamentares nas comissões e no plenário. Em especial, as emendas com temáticas similares são agrupadas e discutidas em conjunto, elevando a complexidade do trabalho de organização realizado pelas equipes técnicas da Câmara dos Deputados Federal e do Senado Federal.

Nessas casas legislativas, a coleta e a organização das emendas são cruciais para otimizar o andamento do trâmite legislativo, entretanto, a quantidade significativa de emendas apresentadas em curtos períodos acarreta sobrecarga de trabalho para as equipes responsáveis, as quais realizam as análises desses documentos de maneira predominantemente manual. A execução manual gera, consequentemente, limitações de agilidade e, ainda mais, de escalabilidade, ponto chave perante um cenário de crescimento constante no volume de projetos colocados em pauta nos plenários federais do Brasil. Dessa forma, a adoção de ferramentas de automação visando acelerar o processo torna-se atrativo e importante, assim, a Inteligência Artificial (IA) apresenta-se como promissora alternativa para lidar com o agrupamento semântico de emendas legislativas.

Nesse contexto, o projeto Ulysses (Câmara dos Deputados, 2019), uma iniciativa em parceria entre o Instituto de Ciências Matemáticas e de Computação (ICMC) da Universidade de São Paulo (USP) e a Câmara dos Deputados Federal, teve por objetivo desenvolver soluções de IA para o contexto legislativo. Este trabalho é fruto desse projeto e busca alinhar o avanço da tecnologia de *Large Language Model* (LLM), modelos de linguagem de grande escala que utilizam aprendizado profundo para processar e entender textos complexos, com a tarefa de agrupamento semântico, podendo explorar as capacidades de execução e desempenho dessa tecnologia em ascendência ao categorizar documentos complexos do legislativo brasileiro.

1.2 Objetivos

O objetivo central deste trabalho é explorar o uso dos modelos GPT-4o e GPT-4o-mini da OpenAI, para realizar o agrupamento semântico de emendas legislativas, permitindo que documentos com temas semelhantes sejam analisados conjuntamente, reduzindo a redundância e tornando a avaliação de mudanças em um PL ou uma PEC mais eficiente. Desse modo, verifica-se a possibilidade de organizar emendas de maneira automatizada e em larga escala, tornando essa abordagem uma alternativa aos métodos tradicionais.

Particularmente, considera-se a capacidade avançada de compreensão contextual e geração de linguagem dos modelos para formar grupos coerentes, podendo, assim, metrificar a execução. Isso faz-se necessário pois, posto que LLMs apresentem alta precisão em diversas tarefas de geração de texto, eles possuem limitações importantes, como a propensão a alucinações — ou seja, a geração de informações incorretas ou sem base no conteúdo fornecido. Dessarte, eles podem apresentar respostas com variabilidade relevante dependendo de parâmetros, como a temperatura, que controla o grau de aleatoriedade nas saídas do modelo. Por isso, investigar como diferentes configurações de temperatura afetam a precisão e a consistência dos agrupamentos, de modo a identificar configurações que ofereçam um bom equilíbrio entre variabilidade e estabilidade nos resultados, também está entre os pontos abordados.

Outro ponto é avaliar o impacto do pré-processamento textual sobre o desempenho dos LLMs na tarefa de agrupamento semântico. Para tal, serão testadas duas abordagens de tratamento textual: passar como *input* os textos das emendas na íntegra, e aplicar pré-processamento a elas. Com isso, pode-se analisar se a simplificação dos textos ajuda o modelo a captar padrões semânticos ou se, ao contrário, prejudica a análise contextual ao remover termos potencialmente relevantes. As métricas de precisão, *recall* e F1 são os critérios de avaliação, que medem a similaridade semântica entre os tópicos gerados e as emendas associadas.

Por fim, pretende-se estabelecer uma base de conhecimento sobre os desafios e limitações do uso de LLMs no contexto legislativo, discutindo as vantagens e as limitações desses modelos e apontando direções para trabalhos futuros. Ao final, espera-se obter uma aplicação capaz de realizar agrupamentos semânticos que sejam uma solução prática para a organização de emendas legislativas, promovendo maior eficiência no trabalho de análise e categorização realizado pelas equipes parlamentares.

1.3 Organização

A presente monografia estrutura-se a partir de quatro capítulos principais, sendo o primeiro este de Introdução. No capítulo 2, apresenta-se a Revisão Bibliográfica, em que é fornecido um panorama sobre os avanços dos LLMs e discute-se o uso de IA no contexto legislativo, destacando algumas abordagens exploradas pelo projeto Ulysses. Em sequência, no

capítulo 3, tem-se o Desenvolvimento, no qual detalha-se todo o processo adotado, incluindo a descrição do *corpus* utilizado, o pré-processamento aplicado às emendas e a metodologia de agrupamento, as configurações de temperatura e as métricas de avaliação utilizadas para validar os resultados. No Capítulo 4, apresenta-se a Conclusão, finalizando com destaques paras as principais contribuições obtidas, analisando as limitações da abordagem implementada e propondo direções para futuros trabalhos que possam aprimorar o uso de LLMs no agrupamento e organização de documentos legislativos.

Capítulo 2

REVISÃO BIBLIOGRÁFICA

2.1 Considerações Iniciais

O uso de IA em aplicações de cunho legislativo expande-se com o tempo, visto que é uma maneira de projetar soluções que superem os desafios impostos por processos manuais somados à complexidade da linguagem legal. Em complemento, os LLMs representam uma possibilidade de serem aplicados de forma a permitir que documentos legais sejam interpretados e categorizados em larga escala.

2.2 O Avanço dos LLMs

2.2.1 Arquitetura Transformer

Sendo a base para o surgimento de LLMs, a arquitetura Transformer, proposta por Vaswani (VASWANI *et al.*, 2017), modela dependências entre as palavras presentes em uma sequência por meio de mecanismos de atenção, permitindo que o modelo processe dependências globais na sequência de entrada. Essa abordagem adotada permitiu maior paralelização em treinamentos e elevou a eficiência para processamentos a partir de sequências longas, manipuladas matematicamente como *tokens* para um processamento mais eficaz. Com isso, os Transformers superaram arquiteturas anteriores de Rede Neural Recorrente (RNN) (JORDAN, 1986) e Rede Neural Convolucional (CNN) (ALBAWI TAREQ ABED MOHAMMED, 2017), obtendo novos resultados de estado da arte em *benchmarks*, permitindo maior escalabilidade ao processar toda a sequência de entrada simultaneamente.

2.2.2 O GPT-4

Desenvolvido pela OpenAI, o modelo GPT-4 apresenta significativas evoluções dentre os modelos baseados em Transformers, produzindo saídas textuais com desempenho superior em *benchmarks* acadêmicos e de iniciativas privadas. Ele foi pré-treinado para prever o próximo *token* em uma sequência e refinado por meio de *Reinforcement Learning from Human Feedback* (RLHF). Ademais, ao se comparar com seu antecessor, o GPT-3.5, esta nova versão mostrou um salto notável, obtendo resultados entre os melhores 10% em simulados de exames da Ordem dos Advogados dos Estados Unidos (OPENAI, 2023).

Apesar de trazer avanços consideráveis, o GPT-4 ainda é afligido pela propensão a alucinações, além de ser limitado pelo tamanho do contexto de memória, ou seja, a quantidade de *tokens* suportados (OPENAI, 2023). Essas características ressaltam a necessidade de cautela na utilização dos resultados do modelo em contextos críticos.

2.2.3 A Estruturação de Prompts

Bsharat, Myrzakhan e Shen (2024) propõe um conjunto de 26 princípios para aprimorar a engenharia de *prompts*, as instruções e comandos passados a um LLM, destacando a importância de um *design* adequado de instruções para que as respostas geradas pelo modelo sejam maximizadas. Para garantir a performance desejada, foram estabelecidos princípios de elaboração de *prompts* eficientes. Constatou-se por meio de um *benchmark* próprio, ATLAS, que para o GPT-4 houve um aumento de 57,7% na qualidade e 36,4% na precisão.

2.3 Trabalhos Relacionados

2.3.1 Uso de LLMs no Domínio Legal

Vayadande *et al.* (2024) explora soluções para o contexto legislativo baseadas em LLMs, em que, além de gerar textos automaticamente, podem também identificar inconsistências e realizar tarefas, tais como análise de sentimentos e Reconhecimento de Entidades Nomeadas (NER), com alta precisão. Utilizando o modelo GPT-3.5 e combinando diferentes técnicas de pré-processamento, normalização dos textos e remoção de *stopwords*, esses sistemas são capazes de detectar anomalias e facilitar a compreensão de extensos e complexos documentos. Em testes, obtiveram maiores taxas de acurácia e menor erro ao comparar com o processo manual e o software sem o uso de LLMs.

2.3.2 Recuperação de Informação no Domínio Legal

Em um contexto brasileiro, no Poder Legislativo há iniciativas voltadas para melhorar a eficiência no acesso e análise de documentos por meio da IA, o projeto Ulysses (SOUZA *et al.*, 2021) reflete isso. Voltado para a Câmara Dos Deputados Federal, existe um *pipeline* de recuperação de informações que visa aumentar a transparência e oferta suporte às atividades legislativas. Para alcançar o objetivo, o sistema desenvolvido utiliza variantes do algoritmo *Best Match 25* (BM25), estas recuperam documentos relevantes em resposta a consultas específicas, tendo foco em propostas legislativas como Projetos de Lei e Propostas de Emendas Constitucional.

Outrossim, a abordagem envolve a remoção de *stopwords*, uso de *stemming* e combinações de unigramas e bigramas para otimizar a relevância dos resultados. Esse *pipeline* é essencial para consultas prévias realizadas pelo corpo técnico da Câmara, que precisa identificar propostas

redundantes ou relacionadas a outras já existentes, evitando esforços duplicados (SOUZA *et al.*, 2021).

2.3.3 Agrupamento de Emendas Legais

2.3.3.1 Senado Brasileiro

Além das iniciativas de recuperação de informação na Câmara dos Deputados, outras frentes relevantes de pesquisa no Brasil têm explorado a organização e análise de emendas legislativas por meio de técnicas avançadas de PLN. No contexto do Senado Federal, Pressato *et al.* (2024) desenvolveu um estudo que avalia diferentes abordagens para agrupar emendas semelhantes com foco na eficiência do processo legislativo.

No estudo, comparou-se as técnicas de BM25, e variações, em relação ao *Sentence Transformer* (SBERT) (REIMERS; GUREVYCH, 2019) no processo de avaliação de similaridade entre emendas a partir de seu conteúdo textual em três cenários: inteiro teor, apenas a seção de Justificação e apenas o Texto Principal.

Como resultados, constatou-se um melhor desempenho, em termos de *recall*, do BM25L, ocorrendo correspondências exatas quando as emendas em inteiro teor foram consideradas. Ademais, o uso do algoritmo de *stemming* Removedor de Sufixos da Língua Portuguesa (RLSP) (MOREIRA; HUYCK, 2001) fez-se essencial para aumentar a precisão dos resultados (PRESSATO *et al.*, 2024).

2.3.3.2 Senado Italiano

Agnoloni *et al.* (2022) explorou técnicas de agrupamento de emendas legislativas do Senado italiano tendo por objetivo a melhora na gestão e aumento de facilidade durante o processo de votação simultânea. No estudo, o desafio principal era identificar grupos de emendas com formulações textuais semelhantes, algumas vezes dispersas ao longo de um dossiê, mas que propunham modificações semelhantes em diferentes partes de uma lei.

O sistema desenvolvido utilizou métricas de similaridade lexical para agrupar textos quase duplicados, priorizando a proximidade textual sobre a similaridade semântica completa. Essa abordagem permitiu ao Senado italiano lidar com o fenômeno de obstrução legislativa, onde partidos da oposição apresentam um grande número de emendas semelhantes com pequenas variações, com o intuito de atrasar o processo legislativo. O algoritmo foi integrado como uma funcionalidade experimental no sistema interno de gestão de emendas, fornecendo uma interface para a detecção, visualização e navegação entre clusters de emendas semelhantes (AGNOLONI *et al.*, 2022).

Esse estudo destaca a importância de extrair e normalizar características específicas do domínio, como citações legislativas dentro dos textos, para melhorar a eficácia do agrupamento automático. Dessarte, a abordagem enfatiza que a criação de *clusters* não requer informações

prévias ou *datasets* manualmente anotados, adotando um modelo não supervisionado para lidar com a variação nos textos legislativos (AGNOLONI *et al.*, 2022).

2.4 Considerações Finais

Em suma, percebe-se que a aplicação de IA no contexto legislativo demonstra grande potencial para otimizar processos, garantir eficiência e melhorar a transparência das atividades parlamentares. Somado a isso, a adoção de LLMs, como o GPT-4, permite que tarefas complexas sejam realizadas de forma automatizada e precisa. Não obstante, a implementação dessas tecnologias também exige cuidado com limitações conhecidas, como a propensão a alucinações e restrições no tamanho do contexto de memória.

Buscou-se, por conseguinte, elaborar um *pipeline* eficiente e seguro que possibilitasse o desenvolvimento de uma aplicação cuja finalidade seja o agrupamento semântico de emendas legislativas por intermédio de um LLM. Para fins de estabelecer o melhor formato, foram avaliadas diferentes abordagens de pré-processamento dos textos de entrada em dois diferentes modelos, sendo assim possível analisar o comportamento de cada um perante alterações no formato de entrada e, também, o valor metrificado do agrupamento feito.

Capítulo 3

DESENVOLVIMENTO

3.1 Considerações Iniciais

Este capítulo detalha em totalidade os aspectos referentes ao projeto de agrupamento semântico de emendas legislativas utilizando LLMs. Inicialmente, descreve-se as emendas legislativas retiradas do Senado Federal Brasileiro e as variações de tratamentos textuais adotados em cada emenda, seguido pela arquitetura do projeto, esta envolve o uso da *Application Programming Interface* (API) da OpenAI para interagir com os modelos GPT-4 e GPT-4 mini. Além disso, explica-se as temperaturas utilizadas nas execuções do modelo, variando de 0.0 a 1.0, e como essa variação impacta a geração dos grupos.

Em seguida, explica-se as atividades realizadas, incluindo o processo de envio das emendas aos modelos com a instrução para agrupar cada emenda em apenas um grupo e gerar um tópico sucinto que resuma o conteúdo do grupo. Visando avaliar os resultados, foi utilizado o BERTScore para calcular as métricas de precisão, *recall* e F1 dos agrupamentos gerados, repetindo os testes cinco vezes por temperatura e por método de pré-processamento adotado para garantir um embasamento estatístico sólido.

Outrossim, apresenta-se os resultados obtidos, com análises sobre o desempenho dos diferentes modelos e temperaturas, bem como aborda-se brevemente as dificuldades e limitações enfrentadas ao longo do processo. Este capítulo finaliza com uma discussão sobre possíveis caminhos alternativos e melhorias que poderiam ser adotadas no futuro.

3.2 Descrição do Projeto

O objetivo deste trabalho é utilizar um LLM para agrupar emendas legislativas do Senado Federal Brasileiro com base em sua similaridade semântica. Cada emenda deve ser atribuída a um único grupo, e o modelo deve fornecer tópicos curtos que descrevam cada grupo gerado.

3.2.1 Corpus

O *dataset* utilizado neste projeto é composto por um total de 6320 emendas legislativas do Senado Federal Brasileiro, abrangendo um período de 13 anos, de 2010 a 2022. As emendas estão distribuídas como consta na Tabela 1:

Ano	Quantidade de Emendas
2010	252
2011	417
2012	949
2013	811
2014	420
2015	415
2016	533
2017	291
2018	461
2019	326
2020	546
2021	522
2022	377

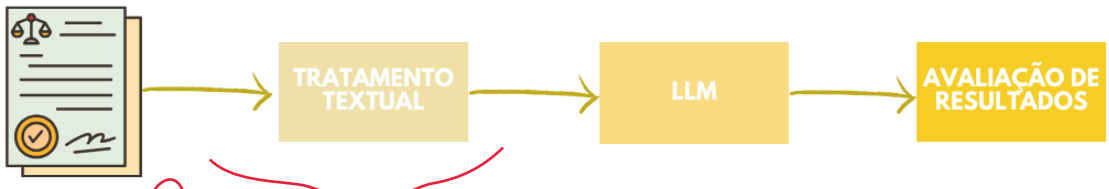
Tabela 1 – Distribuição de Emendas por Ano no *Dataset*

Além de estarem organizadas por ano, as emendas estão agrupadas por projeto legislativo, indicando a qual PL ou PEC se referem. As emendas variam em tamanho e complexidade, contendo tanto teores completos quanto seções explicativas, como justificativas e trechos técnicos. Embora não fosse possível aplicar o programa desenvolvido em todas as emendas, como justificado na Seção 3.5, dispôs-se das emendas referentes ao Art. 21 PL 280 de 2020, 71 documentos, e as referentes ao Art. 4 do PL 8167/2022, 22 documentos.

3.2.2 Metodologia

Para desenvolver o projeto proposto, adotou-se o *pipeline* encontrado na Figura 1.

Figura 1 – Pipeline do Projeto



Fonte: Elaborada pelo autor.

Existe algum tipo de padronização? Nomes de seções?

A fim de medir o impacto de pré-processamento textual no agrupamento, adotou-se duas abordagens diferentes: passar o texto da emenda sem nenhum tratamento; ou retirar *stopwords* e normalizar o conteúdo.

Após o tratamento textual, as emendas são entregues ao LLM, o qual é instruído a realizar um agrupamento semântico com cada grupo possuindo um breve tópico que o descreva e uma emenda podendo pertencer somente a um grupo. Este processo é repetido para cinco valores diferentes de temperatura, 0; 0,25; 0,75; e 1, de forma a mensurar o impacto desse parâmentro

na execução, uma vez que o valor crescente da temperatura, variando de 0 a 1, aumenta o grau de aleatoriedade nas respostas do LLM.

Para validar a qualidade dos agrupamentos, mediu-se a similaridade semântica entre as emendas e os tópicos gerados pelo modelo para cada grupo. A avaliação é realizada com três métricas principais:

Como foi montado o ground-truth?

Precisão: Calcula a proporção de emendas corretamente agrupadas, emendas relevantes, dentro de um grupo, ou seja, quantas emendas atribuídas a um grupo realmente pertencem a ele.

$$\text{Precisão} = \frac{\text{Emendas Relevantes no Grupo}}{\text{Total de Emendas no Grupo}} \tag{3.1}$$

Recall: Avalia a capacidade do modelo de encontrar todas as emendas relevantes para um determinado grupo. Em outras palavras, verifica o quão completo é o agrupamento em relação às emendas que deveriam ser incluídas.

$$\textit{Recall} = \frac{\text{Emendas Relevantes no Grupo}}{\text{Total de Emendas Relevantes Disponíveis}} \tag{3.2}$$

F1-Score: Combina precisão e recall em uma média harmônica, sendo uma métrica que equilibra ambos os aspectos. Um alto valor de *F1-score* indica que o modelo é preciso e completo ao mesmo tempo.

$$\textit{F1-score} = \frac{\text{Precisão} \times \textit{Recall}}{\text{Precisão} + \textit{Recall}} \tag{3.3}$$

Com essas métricas para cada grupo, calcula-se a média de forma a se obter a precisão, o *recall* e o *F1-score* da execução a partir de determinado fluxo e valor de temperatura. Esse processo é repetido 5 vezes, possibilitando uma análise estatística acerca dos resultados.

3.3 Descrição das Atividades Realizadas

Para projetar um sistema capaz de realizar o *pipeline* descrito anteriormente, optou-se por utilizar a linguagem de programação Python devido à sua gama de bibliotecas, em especial àquelas relacionadas com Processamento de Linguagem Natural (PLN).

3.3.1 Tratamento Textual

A partir do arquivo contendo as emendas fornecido como *input*, caso seja aplicado o pré-processamento descrito anteriormente, utiliza-se a biblioteca *Natural Language Toolkit* (NLTK) (BIRD; LOPER, 2004) para a remoção de *stopwords* a partir de uma lista para o português.

3.3.2 Uso de LLMs

O acesso aos modelos GPT-4o e GPT-4o Mini ocorrem por meio da API da OpenAI. Entretanto, a manipulação deles a nível de código de forma a possibilitar que as emendas sejam classificadas, resultando em uma estrutura de dados que engloba cada um dos grupos, seu tópico e as emendas pertencentes a ele, exige o uso de bibliotecas adicionais.

em nível de

A biblioteca LangChain (2023) é projetada para facilitar o desenvolvimento de aplicações que utilizam LLMs. Ela fornece ferramentas para integrar os modelos com outras fontes de dados, estruturar fluxos complexos de conversação, as chamadas *chains* e adicionar lógica de decisão ao processamento das respostas geradas pelos modelos. É eficiente para tarefas que exigem a combinação de LLMs com lógica mais sofisticada, ajudando a manipular o fluxo de informação e controlar interações com modelos de maneira eficiente. Ressalta-se que as cadeias de conversação são as estruturas onde, de fato, são definidas as atividades a serem desempenhadas pelo modelo. Por meio de uma cadeia, estrutura-se o *pipeline* contendo o *prompt* com a instrução, o modelo adotado e informações externas, como perguntas de usuários, no caso em questão, emendas.

No programa, o Langchain fez-se útil ao estruturar uma *chain* capaz de sumarizar o texto de uma emenda passada como *input*. Durante o processo de implementação, foi notória a exigência de um *prompt* rebuscado, sendo necessário elencar algumas restrições para que a integridade do processo de agrupar semanticamente, bem como o formato de saída desejado, fossem respeitados. Isso pode ser visto no Quadro 1.

Função	*Prompt*
Agrupamento Semântico	Você é um assistente de IA especializado no setor legislativo e possui a tarefa de agrupar de forma semântica as emendas parlamentares fornecidas pelo usuário e gerar os grupos contendo o tópico desse grupo e as emendas contidas nele. <Restrictions> 1. SEMPRE gere os grupos em português. 2. Cada emenda deve aparecer SOMENTE em um único grupo. Apenas inclua as emendas, NÃO explique porque ela está no grupo. 3. NÃO pode haver grupo sem emendas nele. 4. NÂO invente emendas, use apenas as passadas a você e os respectivos IDs delas. 5. Gere o tópico de cada grupo de forma sucinta, em no máximo uma frase. 6. A saída deve seguir EXTRITAMENTE o formato: {format_instructions} </Restrictions> As emendas estão abaixo: {amendments}

ESTRITAMENTE

Quadro 1 – *Prompts* Adotados

3.3.3 BERTScore

A avaliação dos agrupamentos semânticos gerados pelos LLMs foi realizada utilizando o BERTScore (ZHANG *et al.*, 2020), uma métrica automática amplamente utilizada para tarefas de avaliação de textos gerados. Valendo-se da similaridade semântica e contextual, difere-se de métricas tradicionais que dependem apenas de correspondências exatas de palavras. É feito um cálculo da pontuação de similaridade entre duas sequências de texto, comparando cada *token* do texto candidato, o tópico do grupo gerado, com cada um do texto de referência, as emendas que pertecem aquele grupo.

Distância do Cosseno para cálculo da similaridade? Não permite indexar em SGBD, pois não é uma métrica (não atende à desigualdade triangular)

Todavia, em vez de realizar uma correspondência exata entre as palavras, são utilizados *embeddings* contextuais – vetores que capturam o significado e o contexto de cada palavra – gerados por um modelo pré-treinado baseado em *Biredctional Encoder Representations from Transformers* (BERT) (DEVLIN *et al.*, 2019). Em sequência, para cada *token* no texto candidato, o BERTScore encontra o mais semelhante no texto de referência, a partir da similaridade do cosseno para mensurar a proximidade entre os vetores de ambos os *tokens*. É possível, pois, captar correspondências entre palavras que, embora diferentes na forma, compartilham significados semelhantes (ZHANG *et al.*, 2020).

3.4 Resultados Obtidos

Nesta seção, apresenta-se o processo de obtenção dos resultados e uma análise detalhada sobre a aplicação dos modelos GPT-4o e GPT-4o-mini no agrupamento semântico das emendas legislativas selecionadas do *corpus*. Ressalta-se que, para cada um dos conjuntos, ambas as abordagens de tratamento textual foram aplicadas para todas as variações de temperatura, repetindo cada um dos processos cinco vezes a fim de angariar dados para executar uma análise estatística dos resultados.

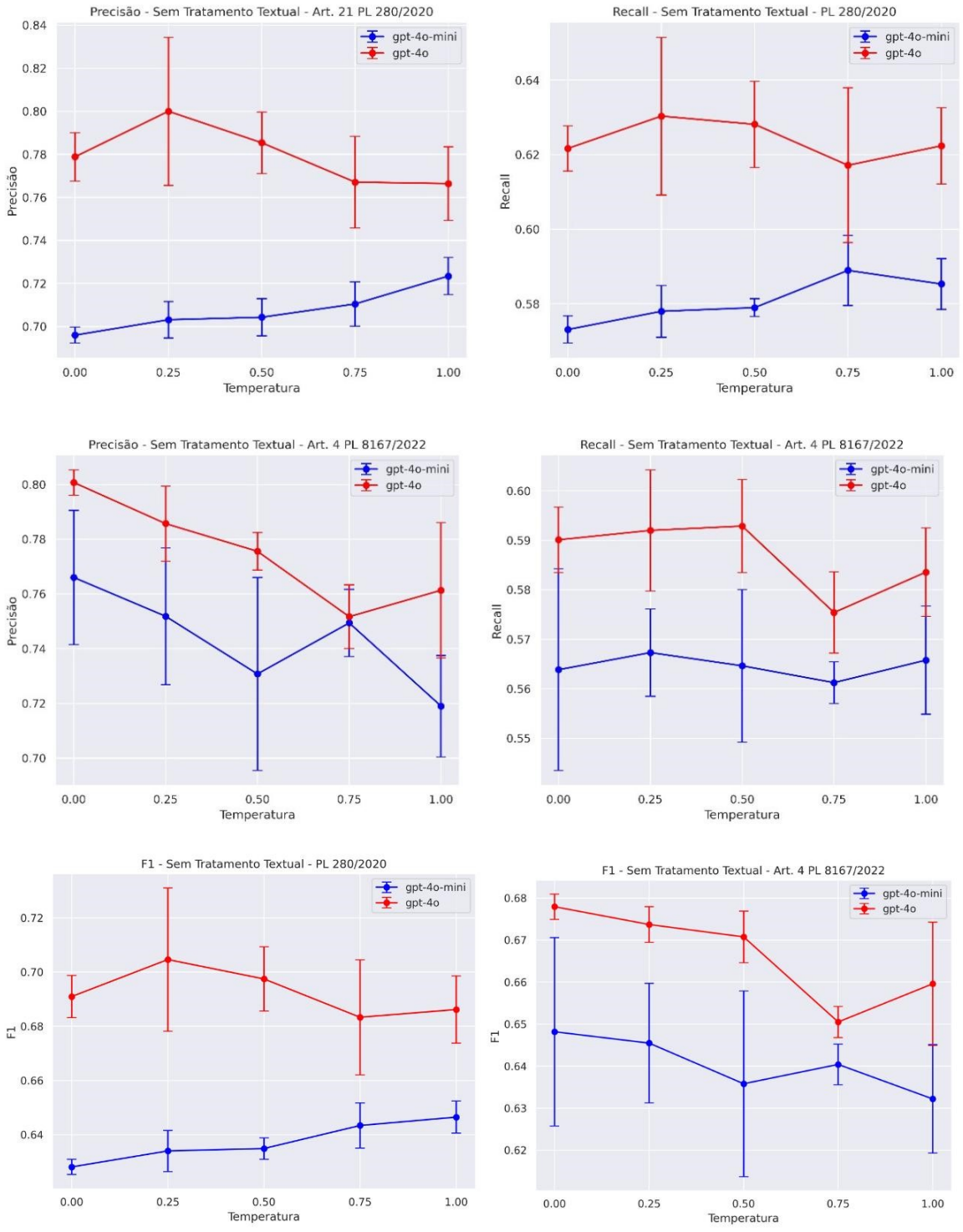
3.4.1 Sem Tratamento Textual

Foi averiguado que o modelo GPT-4o apresentou desempenho superior ao GPT-4o-mini para as métricas adotadas, conforme Figura 2. Pode-se inferir, disso, que a arquitetura mais robusta atrelada à maior capacidade computacional do GPT-4o permitem um processamento mais eficiente dos dados textuais, resultando em uma análise semântica mais precisa. Ressalta-se que, para o GPT-4o, valores de temperatura menores, geraram melhores resultados, percebe-se, disso, que o este modelo é mais otimizado para temperaturas mais controladas.

3.4.2 Com Pré-Processamento

Observou-se, novamente, que o GPT-4o gerou grupos mais precisos e concisos em relação ao GPT-4o-mini ao aplicar a etapa de pré-processamento às emendas, especialmente para

Figura 2 – Gráficos de Métricas para Emendas sem Pré-Processamento



Fonte: Elaborada pelo autor.

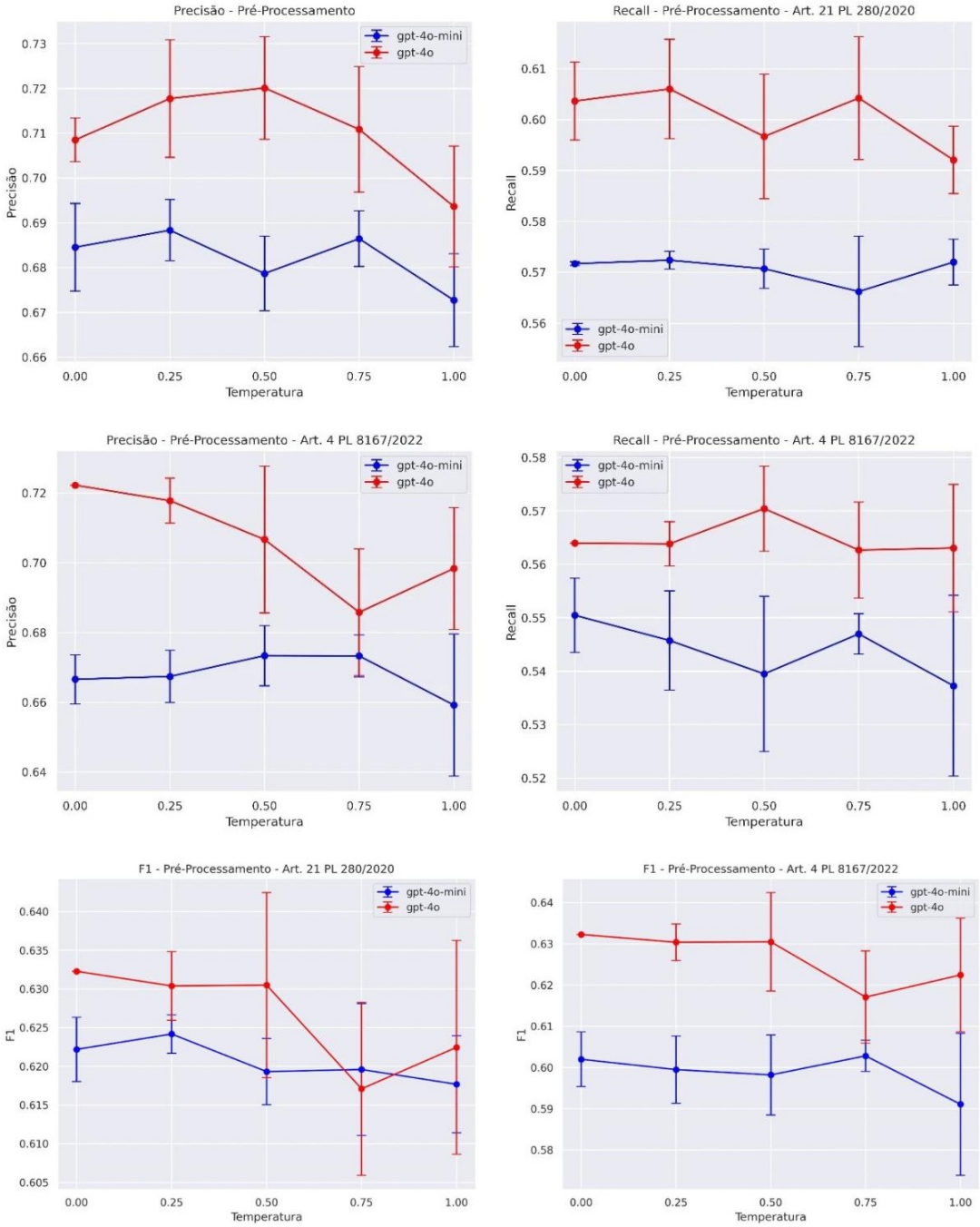
temperaturas menores, conforme Figuras 3. Outrossim, conforme este parâmetro aumentava, o desempenho do primeiro modelo tendeu a diminuir, reforçando que ele propende a gerar agrupamentos menos consistentes conforme a maior tolerância à aleatoriedade nas respostas.

Em geral, os resultados obtidos para as emendas pré-processadas foram menores aos

Poderia indicar em porcentagem, quão menor?

obtidos com emendas sem tratamento textual. Isso evidencia que, embora o pré-processamento simplifique o texto ao remover termos de ligação e outros vocábulos de baixo cunho semântico, o LLM performa melhor ao lidar com o texto completo, conseguindo galgar interpretações mais precisas, formando grupos mais refinados.

Figura 3 – Gráficos de Métricas para Emendas com Pré-Processamento



Fonte: Elaborada pelo autor.

3.5 Dificuldades e Limitações

Embora o objetivo principal do projeto tenha sido alcançado, provando-se possível realizar o agrupamento semântico de emendas legislativas a partir de LLMs, a condução deste trabalho não foi isenta de dificuldades e limitações. Segue, a seguir, as mais impactantes para o resultado final.

3.5.1 Custo de Hardware e API

Inicialmente, pretendia-se adotar LLMs que pudessem ser instalados e executados localmente de maneira totalmente gratuita. Entretanto, para que isso seja feito eficientemente, o *hardware* necessário para abarcar a complexidade e o tamanho desses modelos é robusto, incluindo *Graphic Processing Unit* (GPU) de alto desempenho e grande capacidade de memória, sendo financeiramente inviável para este trabalho. Alternativamente, foi adotado o uso de APIs de um provedor externo, a OpenAI e, apesar de acarretar em um custo menor, ainda foi um fator impactante, sendo a razão pela qual apenas dois conjuntos de emendas foram agrupados e cada agrupamento repetido somente cindo vezes.

3.5.2 Risco de Alucinação

O risco de alucinação, inerente ao uso de LLMs, também teve de ser considerado, visto que, dado o contexto legislativo, é imprescindível que não existam agrupamentos imprecisos. De forma a mitigar tal empecilho, adotou-se *prompts* com restrições que alertavam a não repetir emendas em mais de um grupo, a agrupar todas as emendas e a não usar uma emenda que não estivesse entre as fornecidas, ou seja, não alucinar ao ponto de criar uma emenda inexistente. Além disso, foi feito uma avaliação crítica ao longo do processo para identificar possíveis alucinações.

3.6 Considerações Finais

As análises realizadas permitiram compreender o desempenho dos modelos e a influência do parâmetro de temperatura e da adoção de pré-processamentos nas emendas. Concluiu-se que o GPT-4o se saiu superior ao GPT-4o-mini em métricas de precisão e *recall*, embora esteja longe de alcançar o estado da arte.

O próximo capítulo trará a conclusão deste trabalho, destacando as contribuições realizadas e perspectivas para avanços futuros no uso de LLMs para realizar tarefas envolvendo documentos do campo legislativo, principalmente possibilidades de alcaçar resultados mais satisfatórios.

Capítulo 4

CONCLUSÃO

4.1 Contribuições

Sendo derivado do projeto Ulysses, convênio entre USP e Câmara dos Deputados Federal, este trabalho salienta-se como uma contribuição significativa para o estudo e aplicação de LLMs no contexto legislativo brasileiro. Usufruindo desses modelos para realizarem agrupamentos semânticos de emendas, demonstrou-se a utilidade deles para otimizar e automatizar processos, ao passo que foram evidenciadas as limitações e necessidades de tratamento adequado para alcançar resultados confiáveis e satisfatórios. Apesar de poderosos, os modelos não são soluções prontas e perfeitas, mas sim ferramentas que exigem supervisão e ajustes cuidadosos para que possam ser aplicadas com sucesso.

No âmbito pessoal, a realização deste trabalho, bem como a participação no projeto Ulysses, foram experiências transformadoras, uma vez que possibilitaram-me trabalhar ao lado de colegas de graduação e pós-graduação mais experientes e muito capacitados, logrando-me ensinamentos tanto sobre aspectos técnicos de IA e PLN, quanto sobre a dinâmica de trabalhar em equipe. Ademais, aprendi a importância de documentar metodologias, discutir abordagens e validar resultados com rigor, buscando entender e superar os desafios, composições imprescindíveis para a realização de uma pesquisa científica.

Ao término deste trabalho, percebo-me mais seguro para trilhar o caminho profissional que advém com o encerramento da graduação. As dificuldades encontradas e superadas neste projeto lapidaram minhas capacidades em adaptação e resolução de problemas. Aplicar o método científico, com seu vigor e sua criticidade inerentes, mostrou-se primordial para alcançar resultados de acordo com os objetivos iniciais e trouxe-me uma experiência particular a respeito do papel da ciência em resolver adversidades.

4.2 Considerações Sobre o Curso de Graduação

Visto de um panorama geral, o curso de Engenharia de Computação oferecido pela USP - São Carlos oferece um equilíbrio entre computação e engenharia elétrica, proporcionando ao discente uma bagagem teórica que possibilita transitar e mesclar os saberes dessas duas vertentes.

Toda a experiência que obtive ao longo da graduação contribuiu para que este trabalho

tomasse corpo. Em termos de disciplinas, sobressaem as de Estrutura de Dados - I, II e III - onde foi-me passado os alicerces de programação, e a de Inteligência Artificial, momento chave que despertou meu interesse por essa área e motivou-me a buscar uma iniciação científica nessa temática.

4.3 Trabalhos Futuros

A abordagem adotada mostrou-se capaz de alcançar os objetivos iniciais, visto que os LLMs mostraram-se capazes de executar a tarefa proposta pelos *prompts* passados a eles. Não obstante, as limitações mencionadas na Seção 3.5 ilustram pontos a serem explorados em trabalhos futuros, como a implementação de métodos de *fine-tuning* para adaptação dos modelos ao domínio legislativo brasileiro, de forma a buscar um aprimoramento na coerência das respostas e mitigar o risco de alucinação.

REFERÊNCIAS

AGNOLONI, T.; MARCHETTI, C.; BATTISTONI, R.; BRIOTTI, G. Clustering similar amendments at the Italian senate. In: **Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference**. Marseille, France: European Language Resources Association, 2022. p. 39–46. Disponível em: <https://aclanthology.org/2022.parlaclarin-1.7>. Citado 2 vezes nas páginas 29 e 30.

ALBAWI TAREQ ABED MOHAMMED, S. A.-Z. S. Understanding of a convolutional neural network. In: DEPARTMENT OF COMPUTER ENGINEERING, ISTANBUL KEMERBURGAZ UNIVERSITY, ISTANBUL, TURKEY; ALTINBAS UNIVERSITESI, ISTANBUL, TR; UNIVERSITY OF DIYALA, DIYALA, IQ. **2017 International Conference on Engineering and Technology (ICET)**. Antalya, Turkey: IEEE, 2017. p. 1–6. Date Added to IEEE Xplore: 08 March 2018. Citado na página 27.

BIRD, S.; LOPER, E. NLTK: The natural language toolkit. In: **Proceedings of the ACL Interactive Poster and Demonstration Sessions**. Barcelona, Spain: Association for Computational Linguistics, 2004. p. 214–217. Disponível em: <https://aclanthology.org/P04-3031>. Citado na página 33.

BSHARAT, S. M.; MYRZAKHAN, A.; SHEN, Z. **Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4**. 2024. Acessado em 18 de janeiro de 2024. Citado na página 28.

Câmara dos Deputados. **Câmara lança Ulysses, robô digital para facilitar acesso a informações legislativas**. 2019. [Acessado em: 15 de julho de 2024]. Disponível em: <https://www.camara.leg.br/noticias/548730-camara-lanca-ulysses-robo-digital->. Citado na página 23.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**. Minneapolis, Minnesota, 2019. p. 4171–4186. Disponível em: <https://aclanthology.org/N19-1423>. Citado na página 35.

JORDAN, M. I. **Serial order: a parallel distributed processing approach**. San Diego, California, 1986. Citado na página 27.

LangChain. **LangChain Documentation - Introduction**. 2023. [Acessado em: 12 de janeiro de 2024]. Disponível em: <https://python.langchain.com/docs/introduction/>. Citado na página 34.

MOREIRA, V.; HUYCK, C. A stemming algorithm for the portuguese language. In: **Proceedings of the Eighth International Symposium on String Processing and Information Retrieval (SPIRE 2001)**. Porto, Portugal: IEEE, 2001. Citado na página 29.

OPENAI. **GPT-4 Technical Report**. 2023. Acessado em 4 de março de 2024. Citado 2 vezes nas páginas 27 e 28.

PRESSATO, D.; ANDRADE, P. L. C. de; JUNIOR, F. R.; SIQUEIRA, F. A.; SOUZA, E. P. R.; SILVA, N. F. F. da; DIAS, M. de S.; CARVALHO, A. C. P. de Leon Ferreira de. Natural language processing application in legislative activity: a case study of similar amendments in the Brazilian senate. In: GAMALLO, P.; CLARO, D.; TEIXEIRA, A.; REAL, L.; GARCIA, M.; OLIVEIRA, H. G.; AMARO, R. (Ed.). **Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1**. Santiago de Compostela, Galicia/Spain: Association for Computational Lingustics, 2024. p. 614–619. Disponível em: <https://aclanthology.org/2024.propor-1.70>. Citado na página 29.

REIMERS, N.; GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 2019. Disponível em: <https://arxiv.org/abs/1908.10084>. Citado na página 29.

SOUZA, E.; VITÓRIO, D.; MORIYAMA, G.; SANTOS, L.; MARTINS, L.; SOUZA, M.; FONSECA, M.; FÉLIX, N.; CARVALHO, A. C. P. d. L. F. de; ALBUQUERQUE, H. O.; OLIVEIRA, A. L. I. An information retrieval pipeline for legislative documents from the brazilian chamber of deputies. In: **Legal Knowledge and Information Systems**. [S.l.]: IOS Press, 2021. p. 119–126. Citado 2 vezes nas páginas 28 e 29.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J. Attention is all you need. In: **Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)**. Long Beach, CA, USA: NeurIPS Foundation, 2017. Citado na página 27.

VAYADANDE, K.; BHAT, A.; BACHHAV, P.; BHOYAR, A.; CHAROLIYA, Z.; CHAVAN, A. Ai-powered legal documentation assistant. In: **Proceedings of the 4th International Conference on Pervasive Computing and Social Networking (ICPCSN)**. [S.l.]: IEEE, 2024. ISBN 979-8-3503-8634-9. Citado na página 28.

ZHANG, T.; KISHORE, V.; WU, F.; WEINBERGER, K. Q.; ARTZI, Y. BERTSCORE: Evaluating text generation with BERT. In: CORNELL UNIVERSITY AND ASAPP INC. **Proceedings of the 8th International Conference on Learning Representations (ICLR)**. 2020. Disponível em: <https://openreview.net/pdf?id=SkeHuCVFDr>. Citado na página 35.