UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

HELEN BENTO CAVALCANTI

AVALIAÇÃO DE GRANDES MODELOS DE LINGUAGEM PARA DETECÇÃO DE TÓPICOS E POSICIONAMENTOS EM DEBATES: UM ESTUDO DE CASO NO CONTEXTO DO SENADO FEDERAL.

CAMPINA GRANDE - PB

2024

HELEN BENTO CAVALCANTI

AVALIAÇÃO DE GRANDES MODELOS DE LINGUAGEM PARA DETECÇÃO DE TÓPICOS E POSICIONAMENTOS EM DEBATES: UM ESTUDO DE CASO NO CONTEXTO DO SENADO FEDERAL.

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador : Cláudio Elízio Calazans Campelo

CAMPINA GRANDE - PB

2024

HELEN BENTO CAVALCANTI

AVALIAÇÃO DE GRANDES MODELOS DE LINGUAGEM PARA DETECÇÃO DE TÓPICOS E POSICIONAMENTOS EM DEBATES: UM ESTUDO DE CASO NO CONTEXTO DO SENADO FEDERAL.

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Cláudio Elízio Calazans Campelo

Orientador – UASC/CEEI/UFCG

Leandro Balby Marinho

Examinador – UASC/CEEI/UFCG

Francisco Vilar Brasileiro

Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 15 de Maio de 2024.

CAMPINA GRANDE - PB

RESUMO

O poder legislativo no Brasil é uma das três funções essenciais do Estado. No entanto, há um desafio evidente em relação ao acompanhamento das discussões nos órgãos públicos por parte da população. Isso se deve à extensão considerável e ao volume significativo dessas reuniões, tornando-as inacessíveis para muitos cidadãos. Para enfrentar esse desafio, este estudo utilizou as notas taquigráficas do Senado Federal do ano de 2023, que são transcrições dos debates parlamentares, com o objetivo de avaliar o potencial de Grandes Modelos de Linguagem (do inglês, Large Language Models - LLMs), de detectar tópicos relevantes discutidos pelos parlamentares e o posicionamento deles em relação a esses tópicos, classificando-os como a favor, neutro ou contra. Foram realizados experimentos, ambos utilizando o modelo GPT-3.5-Turbo, para as tarefas mencionadas. O primeiro experimento empregou uma técnica de compressão de dados antes de fornecer a entrada para o GPT e abrangeu reuniões de diferentes tamanhos. O segundo experimento não envolveu compressão e focou apenas em reuniões pequenas. Os resultados indicam que o modelo teve um desempenho superior para reuniões pequenas. Além disso, em um panorama geral para reuniões independentes de tamanho, o modelo teve um desempenho superior na tarefa de detecção de tópicos, com uma precisão média de aproximadamente 70%, enquanto na detecção de posicionamento teve um desempenho razoável com uma precisão média de aproximadamente 60%.

EVALUATION OF LARGE LANGUAGE MODELS FOR DETECTING TOPICS AND STANCES IN DEBATES: A CASE STUDY IN THE CONTEXT OF FEDERAL SENATE.

ABSTRACT

Legislative power in Brazil is one of the three essential functions of the State. However, there is a clear challenge for the population to follow discussions in public bodies. This is due to the considerable length and volume of these meetings, making them inaccessible to many citizens. To address this challenge, this study used the Federal Senate's 2023 tachygraph notes, which are transcripts of parliamentary debates, with the objective of evaluating the potential of Large Language Models (LLMs) to detect relevant topics discussed by parliamentarians and their stance on these topics, classifying them as in for, neutral or against. Experiments were carried out, both using the GPT-3.5-Turbo model, for the tasks mentioned. The first experiment used a data compression technique before providing input to the GPT and covered meetings of different sizes. The second experiment did not involve compression and focused only on small meetings. The results indicate that the model performed better for small meetings. In addition, in a general overview for size-independent meetings, the model performed better in the topic detection task, with an average precision of approximately 70%, while in position detection it performed reasonably well with an average precision of approximately 60%.

Avaliação de grandes modelos de linguagem para detecção de tópicos e posicionamentos em debates: um estudo de caso no contexto do Senado Federal.

Trabalho de Conclusão de Curso

Helen Bento Cavalcanti (Aluno), Cláudio E. C. Campelo (Orientador)
Departamento de Sistemas e Computação
Universidade Federal de Campina Grande
Campina Grande, Paraíba - Brasil

RESUMO

O poder legislativo no Brasil é uma das três funções essenciais do Estado. No entanto, há um desafio evidente em relação ao acompanhamento das discussões nos órgãos públicos por parte da população. Isso se deve à extensão considerável e ao volume significativo dessas reuniões, tornando-as inacessíveis para muitos cidadãos. Para enfrentar esse desafio, este estudo utilizou as notas taquigráficas do Senado Federal do ano de 2023, que são transcrições dos debates parlamentares, com o objetivo de avaliar o potencial de Grandes Modelos de Linguagem (do inglês, *Large Language Models* - LLMs), de detectar tópicos relevantes discutidos pelos parlamentares e o posicionamento deles em relação a esses tópicos, classificando-os como a favor, neutro ou contra. Foram realizados experimentos, ambos utilizando o modelo GPT-3.5-Turbo, para as tarefas mencionadas. O primeiro experimento empregou uma técnica de compressão de dados antes de fornecer a entrada para o GPT e abrangeu reuniões de diferentes tamanhos. O segundo experimento não envolveu compressão e focou apenas em reuniões pequenas. Os resultados indicam que o modelo teve um desempenho superior para reuniões pequenas. Além disso, em um panorama geral para reuniões independentes de tamanho, o modelo teve um desempenho superior na tarefa de detecção de tópicos, com uma precisão média de aproximadamente 70%, enquanto na detecção de posicionamento teve um desempenho razoável com uma precisão média de aproximadamente 60%.

PALAVRAS-CHAVE

Grandes modelos de linguagem, LLMs, Política, Senado Federal, Detecção de Tópicos, Detecção de Posicionamentos.

1 INTRODUÇÃO

O ser humano é inerentemente social, dotado da capacidade de se comunicar, debater e defender seus pontos de vista. Desde os tempos da Grécia Antiga, onde as ágoras eram locais de intensos debates, até os dias atuais, inúmeros cenários proporcionam espaços para discussões e trocas de ideias. Nessas situações, muitas vezes há uma audiência que, embora não participe diretamente das discussões, tem um interesse significativo no que foi debatido. Isso se deve ao fato de que as decisões tomadas nesses debates podem ter um impacto direto ou indireto em suas vidas. Como exemplo é possível citar, assembleias de condomínio, assembleias universitárias, debates parlamentares, entre outros.

O Legislativo Brasileiro representa um cenário peculiar e de grande interesse, uma vez que a população elege seus representantes e tenta acompanhar suas atividades, posicionamentos, debates, gastos e demais ações. Nesse contexto, a computação emerge como uma ferramenta fundamental para auxiliar os cidadãos, por meio de técnicas como mineração, análise e visualização de dados. Diversas iniciativas já utilizam estratégias inovadoras nesse sentido. Um exemplo é o projeto Vidinha de Balada [1], que busca avaliar os gastos dos deputados federais e relacioná-los com sua atuação na Câmara. Outro caso é o blog Empenhados [2], que emprega análise e visualização de dados para disponibilizar informações relevantes à população, como dados sobre transporte público na Paraíba e variações nos preços de produtos em licitações. A plataforma Datapedia Eleições [3] se destaca ao responder questões cruciais sobre a distribuição partidária em cargos eletivos e votos em diferentes regiões.

No entanto, acompanhar os debates nas comissões e as sessões plenárias diárias do legislativo ainda é desafiador para a população, seja pelo grande volume de reuniões ou pela extensão das mesmas. Essa lacuna concede maior liberdade aos representantes, pois a dificuldade de acompanhamento reduz as cobranças por parte da sociedade. Mas esse desafio não é exclusivo desse cenário legislativo, podemos generalizar para outras conjunturas já citadas como assembleias de condomínio, assembleias universitárias, debates parlamentares, entre outras.

Nesse contexto, este trabalho tem como objetivo avaliar o potencial dos LLMs para as tarefas de detecção de tópicos relevantes em debates e de posicionamento dos envolvidos. Para isso, foi delineado um cenário experimental, usando debates do Senado Federal. Esses debates geram notas taquigráficas [4], que compreendem a transcrição das falas registradas durante as sessões plenárias e os encontros das comissões.

Os LLMs são modelos de inteligência artificial desenvolvidos para compreender e gerar linguagem humana, capacitados após extenso treinamento em vastos conjuntos de dados textuais para absorver padrões, contextos e nuances da linguagem natural.

Os experimentos conduzidos em nosso estudo visam responder as seguintes questões de pesquisa:

[1] https://www.jusbrasil.com.br/noticias/hackfest-cgu-incentiva-uso-da-tecnologia-para-combate-a-corrupcao-e-exercicio-da-cidadania/479985700
[2] https://analytics-ufcg.github.io/empenhados/
[3] https://eleicoes.datapedia.info/
[4] https://www12.senado.leg.br/noticias/materias/2011/07/22/notas-taquigraficas-permitem-ao-cidadao-acesso-rapido-ao-trabalho-do-senado

- **Q1** - Qual a eficácia de LLMs, mais especificamente do modelo GPT-3.5 Turbo, na tarefa de detecção de tópicos relevantes, no contexto de debates do Senado Federal?
- **Q2** - Qual a eficácia de LLMs, mais especificamente do modelo GPT-3.5 Turbo, na tarefa de detecção de posicionamentos, no contexto de debates do Senado Federal?
- **Q3** - A compressão dos dados de entrada do modelo impactam na eficácia dos resultados?

Além disso, este trabalho acrescenta três contribuições principais para ciência da computação: (1) Um *prompt* generalizável, resultante de um intenso processo de engenharia de *prompt*, para ser aplicado nas tarefas de detecção de tópicos e posicionamentos em cenários de debates como um todo, onde é possível parametrizar e escolher o contexto a ser aplicado; (2) Uma avaliação dessa abordagem, investigando da efetividade do uso de LLMs para as tarefas citadas no contexto das reuniões de 2023 do Senado Federal; (3) E uma base de dados [2] relevante para trabalhos futuros, sendo composta pelas notas taquigráficas das reuniões do Senado Federal de 2023, contendo anotações manuais para possibilitar a avaliação da performance de modelos nas tarefas de interesse;

2 TRABALHOS RELACIONADOS

O aumento no número de pesquisas relacionadas a grandes modelos de linguagem em uma variedade de campos é inegável. Um exemplo ocorre no campo da recomendação de notícias, onde identificar se dois artigos expressam o mesmo ponto de vista é uma estratégia relevante. Uma abordagem crucial para essa determinação é a detecção de posicionamento. Nesse contexto, pesquisadores em um estudo recente [6] testaram duas tarefas (*Same Side Stance* e *Pro/Con*) utilizando duas arquiteturas de Grandes Modelos de Linguagem (*bi-encoding* e *cross-encoding*), além de incorporar conhecimentos relacionados à tarefa (*Natural Language Inference knowledge*, NLI). Suas conclusões destacam que um sistema de recomendação de notícias pode se beneficiar de modelos de posicionamento robustos que consideram uma gama diversificada de tópicos, permitindo a medição eficaz das diferenças de posicionamento entre os artigos. Isso, por sua vez, capacita o sistema a oferecer aos leitores novas perspectivas, assim como na atual pesquisa, usamos LLMs justamente para ajudar, a pessoas interessadas, no entendimento e interpretação de um determinado debate.

No contexto de mídias sociais, um estudo recente [8], apresenta experimentos com LLMs para detecção de posicionamentos em cenários de aprendizado *zero-shot*[5] e *few-shot learning*[6]. Os resultados reportados indicam que LLMs demonstram desempenho superior aos *baselines* nessa tarefa, com LLaMa-2 e Mistral-7B mostrando eficiência e potencial apesar de seus tamanhos menores em comparação com o ChatGPT. Fazendo um paralelo com este estudo, ambos destacam o desempenho dos LLMs em suas respectivas tarefas, com o primeiro estudo enfatizando a eficácia na detecção de posicionamentos no domínio de mídias sociais e o estudo atual demonstrando a capacidade de detecção não só posicionamentos, mas também os tópicos que estão sendo discutidos.

[5]É uma abordagem de aprendizado de máquina onde um modelo é capaz de realizar tarefas para as quais não foi explicitamente treinado, sem fornecer exemplos na fase de inferência.

[6]Técnica que permite que um modelo faça previsões para novas classes a partir de poucos exemplos na fase de inferência.

Além disso, os LLMs estão sendo aplicados na área de *Argument-Mining* (mineração de argumentos), que consiste na extração e identificação automática de estruturas argumentativas em textos em linguagem natural com o auxílio de ferramentas computacionais. Um estudo relacionado a essa área [5] utilizou o modelo GPT-4 para extrair argumentos de *podcasts*, que surgiram como uma plataforma significativa para a troca de ideias, debates, opiniões e conhecimento sobre uma variedade de tópicos. Essa abordagem se relaciona com o presente trabalho devido ao contexto semelhante, no qual a troca de ideias e exposição de opiniões são frequentes. Além de utilizar um LLM para realizar essa tarefa.

No contexto da análise de conjuntos de dados textuais no cenário político, as técnicas de Processamento de Linguagem Natural têm desempenhado um papel significativo. Um estudo brasileiro recente [3] avalia o uso individual e combinado de duas técnicas de ponta para a modelagem de tópicos latentes e a estimativa de pontos ideais com base em texto, aplicando-as para caracterizar os discursos e posicionamentos políticos de parlamentares brasileiros. Especificamente, foram empregados os modelos BERTopic e Text-Based Ideal Point para analisar a 55ª e 56ª Legislaturas da Câmara dos Deputados, no período de 2015 a 2022. Os resultados obtidos e as análises decorrentes indicam a viabilidade e a promessa dessas técnicas para fundamentar novos estudos políticos no contexto brasileiro.

Em outro estudo, técnicas de extração não supervisionada de posicionamento, modelagem de tópicos e rotulagem automatizada foram empregadas [7]. Utilizando postagens retuitadas como interações dos usuários sobre o tema CPI da Covid-19, foram calculadas semelhanças entre os mais ativos em uma discussão. A detecção de posicionamento foi realizada por meio de técnicas de redução de dimensionalidade e clusterização, modelagem de tópicos com *embeddings* contextualizados e rotulagem automática de *clusters* com base em termos recorrentes em cada grupo. Essa abordagem resultou em um pequeno número de *clusters* de usuários (entre 2 e 3), com uma uniformidade na rotulagem dos usuários em diferentes conjuntos superior a 98%.

Este estudo diferencia-se dos trabalhos citados, principalmente porque visa não apenas detectar tópicos, mas também identificar os posicionamentos dos debatedores em relação a esses tópicos, considerando os diversos contextos políticos. Além disso, um fator distintivo em relação a todos os demais estudos mencionados é a utilização de técnicas de compressão de *prompt* de ponta antes da aplicação de um LLM.

3 METODOLOGIA

Esta seção apresenta a metodologia adotada para avaliar o desempenho do modelo nas tarefas desejadas. Discute-se como foi construída a base de dados, os modelos utilizados, e os experimentos conduzidos.

3.1 Extração e pré-processamento de dados

Neste trabalho, são analisados tópicos relevantes em reuniões parlamentares e os posicionamentos dos participantes em relação a esses tópicos. As reuniões são transcritas e transformadas em dados textuais.

Foi criada uma base de dados focalizada nas sessões do Senado Federal[7] de 2023, sem distinguir entre sessões plenárias e reuniões de comissões. Essa base foi criada com o intuito de centralizar as informações das reuniões de interesse durante o período de tempo desejado. Para isso, utilizou-se a página web oficial do Senado Federal como fonte de dados, realizando-se uma raspagem de dados da página. O objetivo principal dessa raspagem foi obter as Notas Taquigráficas, que, de acordo com o site da Câmara dos Deputados[8], "são o conjunto de discursos registrados nas sessões plenárias e nas reuniões das comissões. Um discurso é a transcrição individual de um pronunciamento feito por um parlamentar ou não parlamentar."

Todo o ferramental para extração e pré-processamento de dados foi implementado na linguagem Python, utilizando-se as bibliotecas Pandas[9] e BeautifulSoup[10], e está disponível em um repositório público no GitHub [11]. Além disso, a base de dados completa produzida foi disponibilizada publicamente para download [2].

A extração de dados foi realizada da seguinte forma: primeiramente, extraíram-se todas as *URLs* que direcionam para as notas taquigráficas das reuniões de 2023. Posteriormente, para extrair o conteúdo em si, acessou-se cada *URL* e extraiu-se o identificador único da sessão, o nome da pessoa que proferiu o discurso, o partido da pessoa e o discurso em si. A Tabela 1 apresenta a estrutura do conjunto de dados produzido.

Tabela 1: Estrutura do conjunto de dados sobre eventos realizados no Senado Federal (2023).

Atributo	Descrição
id session	Identificador único do evento
speaker name	Nome da pessoa que proferiu o discurso
party	Partido político
speech	Discurso

Durante o pré-processamento, foi necessário utilizar a *tag* "INDEFINIDO" para alguns casos em que o partido não era explicitado. Em outro caso, quando se tratava do discurso do presidente, a *tag* correspondente ao nome era apenas "PRESIDENTE"; dessa forma, substituiu-se pela pessoa em questão. Posteriormente, a sessão com o ID 25779 teve que ser removida devido à indisponibilidade das notas dessa reunião no site do Senado.

3.2 Compressão do *prompt*

Os modelos de linguagem não interpretam o texto apenas como uma sequência de palavras; ao invés disso, eles o veem como uma sequência de números, conhecidos como *tokens*. A codificação de pares de *bytes* (*Byte Pair Encoding*, BPE [1]) é uma técnica utilizada para transformar texto em *tokens*.

Quando fornecemos o texto de entrada ao modelo, ele o converte em *tokens*. No entanto, é importante observar que cada modelo possui sua própria janela de contexto, que representa a quantidade máxima de *tokens* aceitos como entrada. Em certos cenários, o contexto pode ser muito extenso para ser completamente compreendido pela janela de contexto do modelo. É aqui que surgem as estratégias de compressão de *prompt*, que, além de resolver essa limitação, também podem reduzir custos e a latência.

Existem várias abordagens para comprimir os *prompts*. A utilizada neste estudo emprega modelos de linguagem menores, treinados para identificar *tokens* que não são considerados essenciais [4]. Isso permite que os modelos continuem a entender as entradas, mesmo após a compressão.

O modelo GPT-3.5-Turbo possui uma janela de contexto limitada a 16.385 *tokens* e não aceita entradas maiores. Conforme observado na Figura 1, apenas 41 das 203 reuniões contêm até 16.000 *tokens*, quantidade necessária considerando uma pequena variação no número de *tokens* da saída da compressão.

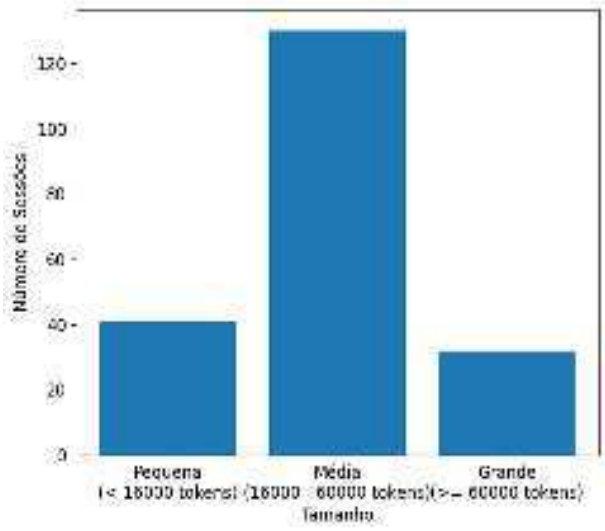

Figura 1: Distribuição da quantidade de sessões por tamanho

Para contornar esse desafio, adotou-se uma estratégia de compressão de dados utilizando o LongLLMLingua [4] , um framework desenvolvido pela Microsoft que facilita a compressão rápida em contextos extensos. O LongLLMLingua[12] é uma estratégia de compressão inovadora que pode obter melhor desempenho em diferentes tarefas em comparação com o *prompt* original. Em geral, o *prompt* comprimido pode alcançar um melhor desempenho com menos custo.

Na implementação dessa estratégia, foram definidos parâmetros essenciais. O principal é o *target token*, que representa o limite máximo de *tokens* que o modelo de compressão deve alcançar, estabelecido em 16.000 para otimizar a performance, visto que a eficácia da compressão decai com o aumento da quantidade de dados comprimidos. Outros dois parâmetros significativos são *context* e *question*: *context* refere-se ao contexto adicional necessário para responder a perguntas baseadas nos diálogos das reuniões, enquanto *question* diz respeito às instruções fornecidas ao modelo de linguagem, como pedidos de informação ou questões específicas. A configuração final desses parâmetros está detalhada na Seção 3.3.

[7] https://www12.senado.leg.br/hpsenado
[8] https://www.camara.leg.br/
[9] https://pandas.pydata.org/
[10] https://beautiful-soup-4.readthedocs.io/en/latest/#
[11] https://github.com/helenbc/tcc-notas-taquigraficas/tree/main/web_scraping
[12] https://github.com/microsoft/LLMLingua

3.3 Engenharia de *prompt*

O *prompt* proposto foi produzido a partir de um meticuloso processo de engenharia de *prompt*, seguindo-se a metodologia proposta pela Deeplearning.ai [13] em parceria com a OpenAI, empresa responsável pelo desenvolvimento dos modelos da família GPT. Durante o processo de engenharia de *prompt*, diversas versões preliminares foram avaliadas. A Figura 2 exibe a versão final obtida para o *prompt*.

Pode-se dividir o *prompt* em três partes:

- Na primeira parte, solicita-se que o modelo se comporte como um especialista em detecção de posicionamento. Define-se o conceito de detecção de posicionamento e explica-se o significado de ser a favor, contra ou neutro em relação a um tópico.
- Em seguida, utiliza-se a segunda tática apresentada no curso, que consiste em solicitar uma saída estruturada, neste caso, um JSON, especificando as chaves e os valores relacionados a cada chave.
- Na segunda parte, explica-se como é o texto de entrada, utilizando-se a primeira tática do curso, que envolve o uso de delimitadores, como “‘ (três acentos graves) para indicar claramente as partes distintas da entrada. Nesse caso, indica-se uma variável chamada *context*, tornando o *prompt* mais generalizado para uso em outros contextos.
- Na terceira e última parte, também seguindo as orientações do curso, estrutura-se um esquema de *pipeline* com uma sequência de pontos de ação para o modelo. Isso é feito com o objetivo de garantir que o modelo execute todas as tarefas esperadas. Além disso, nessa parte, fornece-se ao modelo uma lista das pessoas que fizeram algum pronunciamento no debate.

3.4 Definição dos experimentos

Foram realizados dois experimentos. O primeiro experimento foi delineado para responder as questões de pesquisa **Q1** e **Q2**, que tem o interesse de medir a eficácia do modelo GPT-3.5 Turbo, nas tarefa de detecção de tópicos e posicionamentos, no contexto de debates do Senado Federal. Por sua vez, o segundo experimento está alinhado com a **Q3**, buscando entender se compressão dos dados de entrada impactam na eficácia dos resultados do modelo.

O primeiro experimento, como mostra a Figura 3, é dividido em duas etapas: a compressão do texto e a inferência utilizando o modelo GPT-3.5-Turbo. Na primeira etapa, cada reunião é inserida como *context* no modelo de compressão. O texto da reunião é formatado com falas dispostas da seguinte maneira: *"nome" : "fala"*, onde o nome é marcado com a *tag <llmlingua, compress=False>*, para prevenir a compressão desse segmento crucial na detecção do posicionamento da pessoa. O parâmetro *question* corresponde ao *prompt* descrito na Subseção 3.3, e o *target token* é de 16.000, permitindo uma leve variação na quantidade de *tokens* de saída.

[13] https://www.deeplearning.ai/short-courses/chatgpt-prompt-engineering-for-developers/

Consider that it is an expert model in Stance Detection. Stance detection is the task of predicting an author's point of view on a subject of interest. A speech can represent one of four types of stance: *for, against or neutral.*

For: When an author takes a stance "for" a subject, it means they support or advocate for it. Their speech or writing will likely include arguments, evidence, or opinions that highlight the positive aspects, benefits, or reasons to endorse the subject. For example, if the subject is a proposed policy change, someone taking a "for" stance might emphasise how it could improve people's lives or address important societal issues.
Against: This stance indicates opposition or disagreement with the subject at hand. Authors taking an "against" stance will present arguments, evidence, or opinions that highlight flaws, risks, negative consequences, or reasons to reject the subject. Using the previous example of a proposed policy change, someone taking an "against" stance might argue that it would be ineffective, unfair, or harmful to certain groups.
Neutral: A neutral stance means the author does not express explicit support or opposition towards the subject. They may present information, analysis, or perspectives in a balanced and objective manner without advocating for or against the subject. Neutral stances typically avoid strong opinions or judgements and instead focus on providing a comprehensive understanding of the topic without bias. If the person doesn't say anything about that topic, it means that they should not be listed.

Reply in JSON format with the following keys: list_latent_topics, stances and summary.

- **list_latent_topics:** should contain the list of all topics discussed in the text, and a short description for each topic.
- **stances:** for each latent_topics key should contain the list of classification of the related speaker's speeches.
- **summary:** should contain the summary of the text. Consider that you will receive as input a text with a set of speeches that make up **“‘{context}”’**.

Do the following actions for the text:

- Determine the all topics being discussed in the text and a brief descriptions of these topics.
- For each topic and for each speaker, except if the person doesn't say anything about that topic, classify the stance as being *FOR, AGAINST, NEUTRAL*. Being the following speakers: **{speakers_string}**.
- Before your response, translate the summary and the topics to Portuguese.

Figura 2: Versão Final do Prompt

Figura 3: Diagrama de passos do primeiro experimento

Este processo resulta em um JSON contendo a compressão de cada reunião, que é então convertido em *string* e fornecido como entrada para o GPT, que retorna, também em formato JSON, os tópicos relevantes da reunião e o posicionamento de cada pessoa em relação a esses tópicos.

O segundo experimento, como mostra a Figura 4, consiste em submeter diretamente ao modelo GPT-3.5-Turbo as reuniões menores, com até 15.000 *tokens*, sem compressão, esse número foi usado devido ao espaço de *tokens* necessários para o *prompt*. A estrutura do retorno é a mesma do experimento anterior.

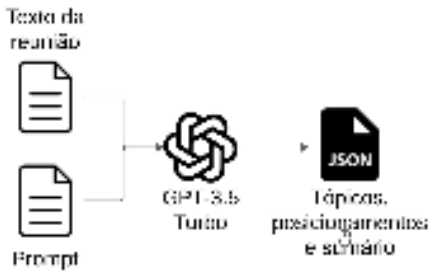

Figura 4: Diagrama de passos do segundo experimento

3.5 Avaliação do modelo

A avaliação do modelo foi dividida em duas partes. A primeira parte refere-se aos tópicos retornados, enquanto a segunda trata do posicionamento das pessoas em relação a esses tópicos.

Na primeira parte da avaliação, são utilizadas as definições de precisão, revocação e F1 de recuperação da informação. Precisão é a fração de tópicos recuperados que são relevantes, revocação é a fração de tópicos relevantes que são recuperados, e a F1 é a média harmônica dessas duas métricas.

Assim, como pode ser visto na Tabela 2, um verdadeiro positivo é um tópico relevante que foi recuperado, um verdadeiro negativo não é relevante e não foi recuperado, um falso positivo não era relevante mas foi recuperado, e um falso negativo é relevante mas não foi recuperado. As fórmulas de acurácia, precisão, revocação e F1 podem ser visualizadas nas Equações 1, 2, 3 e 4.

Tabela 2: Matriz de Confusão para Avaliação de Resultados

	Relevante	Não Relevante
Recuperado	Verdadeiro Positivo (VP)	Falso Positivo (FP)
Não Recuperado	Falso Negativo (FN)	Verdadeiro Negativo (VN)

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \tag{1}$$

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{2}$$

$$\text{Revocação} = \frac{VP}{VP + FN} \tag{3}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{4}$$

Na segunda parte da avaliação, para medir o desempenho do modelo em detectar os posicionamentos, utilizamos a acurácia, precisão, revocação e F1-score da biblioteca Scikit-learn [14]. Como se trata de um problema multiclasse, mapeamos os posicionamentos **a favor** como 0, **neutro** como 1 e **contra** como 2. Dessa forma, para cada tópico predito corretamente pelo modelo, coletamos os posicionamentos de cada parlamentar que participou da reunião e comparamos com as anotações manuais, calculando as métricas mencionadas anteriormente.

4 RESULTADOS E DISCUSSÕES

Esta seção apresenta e analisa os resultados dos experimentos realizados. Serão respondidas as questões de pesquisa:

- **Q1** - Qual a eficácia de LLMs, mais especificamente do modelo GPT-3.5 Turbo, na tarefa de detecção de tópicos relevantes, no contexto de debates do Senado Federal?
- **Q2** - Qual a eficácia de LLMs, mais especificamente do modelo GPT-3.5 Turbo, na tarefa de detecção de posicionamentos, no contexto de debates do Senado Federal?
- **Q3** - A compressão dos dados de entrada do modelo impactam na eficácia dos resultados?

Os experimentos foram realizados em uma máquina equipada com uma GPU RTX 2080 Ti, que possui 11 GB de memória de GPU. Esse ambiente foi especialmente crucial para o primeiro experimento, envolvendo a compressão dos dados. Portanto, optamos pelo modelo quantizado Llama-2-7b-Chat-GPTQ da biblioteca LLMLingua, o qual requer no mínimo 8 GB de memória de GPU.

Os cálculos das métricas foram realizados com base em um conjunto de dados de referência composto por seis reuniões, distribuídas em dois grupos de cada tamanho (pequeno, médio e grande), totalizando 1013 minutos de reunião. Essas reuniões foram manualmente anotadas em um JSON seguindo a estrutura semelhante a estrutura de saída no modelo. É importante observar que, idealmente, teríamos realizado uma amostragem estratificada para garantir a representatividade proporcional dos diferentes tipos de reuniões nos dados reais. No entanto, devido à disponibilidade limitada de recursos humanos para esta tarefa, isso não foi possível.

4.1 Q1 - Qual a eficácia de LLMs, mais especificamente do modelo GPT-3.5 Turbo, na tarefa de detecção de tópicos relevantes, no contexto de debates do Senado Federal

Em relação a primeira questão de pesquisa e consequentemente ao primeiro experimento, foi identificado que o modelo apresenta um desempenho superior na detecção de tópicos em reuniões de pequeno porte, conforme demonstrado na Tabela 3, onde todas as métricas indicaram resultados superiores para esse tipo de reunião. Adicionalmente, observou-se que à medida que o tamanho das reuniões aumenta, ocorre uma queda nas métricas, especialmente em reuniões maiores, onde essas métricas sofrem uma redução significativa.

[14] https://scikit-learn.org/stable/index.html

É importante notar que, em todos os casos, a precisão foi superior à revocação, chegando a 100% para as reuniões pequenas avaliadas. Isso sugere que o modelo tende a identificar corretamente os tópicos mencionados na reunião, com uma baixa probabilidade de erroneamente atribuir um tópico que não foi abordado. Destaca-se que essa alta precisão é especialmente notável em reuniões pequenas e médias.

Quanto à revocação, observou-se uma métrica satisfatória apenas para reuniões pequenas, enquanto para as médias e grandes, esse valor fica abaixo de 50%. Isso indica que o modelo não consegue recuperar todos os tópicos relevantes discutidos durante a reunião nessas situações de maior porte.

Estes resultados ressaltam a importância de desenvolver estratégias específicas para lidar com diferentes tamanhos de sessão. Enquanto o modelo pode ser eficaz em reuniões pequenas, pode ser necessário um ajuste ou uma abordagem diferente para melhorar o desempenho em reuniões médias e grandes.

Após uma análise qualitativa, observou-se que o modelo apresenta uma tendência de extrair com mais precisão os tópicos discutidos no início e no final da reunião, enquanto tende a negligenciar os tópicos abordados no meio.

Tabela 3: Média das Métricas de Desempenho para Detecção de Tópicos por Tipo de Sessão

Tipo de Sessão	Precisão	Revocação	F1	Acurácia
Pequenas	**1.0000**	**0.7500**	**0.8333**	**0.7500**
Médias	0.7500	0.4804	0.5130	0.3471
Grandes	0.3750	0.0551	0.0959	0.0505
Média geral	**0.7083**	**0.4285**	**0.4807**	**0.3825**

4.2 Q2 - Qual a eficácia de LLMs, mais especificamente do modelo GPT-3.5 Turbo, na tarefa de detecção de posicionamentos, no contexto de debates do Senado Federal

Ainda observando o primeiro experimento, que também está relacionado a segunda questão de pesquisa, conforme mencionado na Subseção 3.5, a detecção de posicionamentos é abordada como um problema multiclasse. Posicionamentos a favor, neutro e contra são codificados como 0, 1 e 2, respectivamente. Ademais, membros presentes em reuniões que não intervieram em um debate específico, assim como aqueles que discursaram sem definir claramente seu apoio ou oposição, foram classificados como neutros.

Na tarefa de detecção de posicionamentos, como mostrado na Tabela 4, observam-se métricas inferiores em relação à detecção de tópicos. De modo similar à tarefa anterior, o modelo teve um desempenho superior em reuniões menores, alcançando uma acurácia próxima de 90%. Importante destacar que o desempenho diminui à medida que aumenta o tamanho da reunião, um padrão recorrente nessa tarefa. Contudo, em reuniões de grande porte, o desempenho melhorou significativamente em comparação com a detecção de tópicos.

Interessantemente, as sessões de tamanho médio apresentaram um desempenho inferior nesta tarefa, assemelhando-se mais às métricas observadas em reuniões de grande porte do que às de pequeno porte.

Uma análise mais qualitativa da resposta do modelo revela que, principalmente em reuniões de médio e grande porte, o modelo tende a categorizar todos os participantes da reunião, mesmo aqueles que não se pronunciaram sobre o assunto, com um único posicionamento. Esta classificação parece ocorrer sem seguir uma lógica discernível, como, por exemplo, a simples atribuição da opinião da maioria a todos os presentes. Isso fica evidente em dois exemplos da reunião 25790.

No primeiro caso, referente ao tópico "Projeto de Lei nº 2.788, de 2019", enquanto os dados reais indicam um posicionamento a favor por parte de determinados membros, a saída do modelo classifica a maioria dos participantes como neutros, incluindo até mesmo aqueles que expressaram sua posição sobre o assunto, conforme mostrado na Figura 5.

```
Dados reais:
"Projeto de Lei nº 2.788, de 2019": {
    "A SRA. DAMARES ALVES": "FOR",
    "O SR. JORGE KAJURU": "FOR",
    "A SRA. ZENAIDE MAIA": "FOR"
},
Saída do modelo:
"Projeto de Lei nº 2.788, de 2019": {
    "A SRA. DAMARES ALVES": "NEUTRAL",
    "O SR. PRESIDENTE RODRIGO PACHECO": "NEUTRAL",
    "O SR. EDUARDO GIRÃO": "NEUTRAL",
    "O SR. FLÁVIO ARNS": "NEUTRAL",
    "O SR. CONFÚCIO MOURA": "NEUTRAL",
    "O SR. ORIOVISTO GUIMARÃES": "NEUTRAL",
    "O SR. JAQUES WAGNER": "NEUTRAL",
    "O SR. ESPERIDIÃO AMIN": "NEUTRAL",
    "O SR. ZEQUINHA MARINHO": "NEUTRAL",
    "A SRA. SORAYA THRONICKE": "NEUTRAL",
    ...
    "O SR. MARCELO CASTRO": "NEUTRAL"
}
```

Figura 5: Exemplo de comparativo dos dados reais com a saída do modelo onde a maioria dos posicionamentos é neutro

No segundo exemplo, relacionado ao "Substitutivo da Câmara dos Deputados ao Projeto de Lei nº 4.727, de 2020", novamente os dados reais indicam um posicionamento a favor de certos membros, porém o modelo os rotula como contra, como evidenciado na Figura 6. Esse acaba sendo um erro mais grotesco em relação ao anterior, porque além de rotular com um posicionamento contrario pessoas que não se pronunciaram sobre o tema, o modelo fez o mesmo com pessoas que se pronunciaram claramente a favor daquele tópico.

```
Dados reais:
"Substitutivo da [...] nº 4.727, de 2020": {
    "O SR. PRESIDENTE RODRIGO PACHECO": "FOR",
    "A SRA. SORAYA THRONICKE": "FOR"
},

Saída do modelo:
"Substitutivo da [...] Lei nº 4.727, de 2020": {
    "A SRA. DAMARES ALVES": "AGAINST",
    "O SR. PRESIDENTE RODRIGO PACHECO": "AGAINST",
    "O SR. EDUARDO GIRÃO": "AGAINST",
    "O SR. FLÁVIO ARNS": "AGAINST",
    "O SR. CONFÚCIO MOURA": "AGAINST",
    "O SR. ORIOVISTO GUIMARÃES": "AGAINST",
    "O SR. JAQUES WAGNER": "AGAINST",
    "O SR. ESPERIDIÃO AMIN": "AGAINST",
    "O SR. ZEQUINHA MARINHO": "AGAINST",
    "A SRA. SORAYA THRONICKE": "AGAINST",
    ...
    "O SR. MARCELO CASTRO": "AGAINST"
}
```

Figura 6: Exemplo de comparativo dos dados reais com a saída do modelo onde a maioria dos posicionamentos é contra

Tabela 4: Média das Métricas de Desempenho para Detecção de Posicionamento por Tipo de Sessão

Tipo da Sessão	Precisão	Revocação	F1	Acurácia
Pequenas	**0.8125**	**0.8750**	**0.8333**	**0.8750**
Médias	0.5030	0.5918	0.5251	0.5918
Grandes	0.4640	0.4464	0.4541	0.4464
Média Geral	**0.5932**	**0.6377**	**0.6042**	**0.6377**

4.3 Q3 - A compressão dos dados de entrada do modelo impactam na eficácia dos resultados

Ao analisar o segundo experimento, que está diretamente ligado a terceira questão de pesquisa, constatou-se que a compressão dos dados não teve nenhum impacto no desempenho do modelo na detecção de tópicos em sessões pequenas, como é possível observar na Tabela 5. Idealmente, seria proveitoso conduzir testes adicionais com conjuntos de dados mais extensos e reuniões de tamanhos diversos, aplicando diferentes níveis de compressão. Isso permitiria uma avaliação mais abrangente e confiável para afirmar com segurança que a compressão não afeta os resultados.

Tabela 5: Média das Métricas de Desempenho para Detecção de Tópicos para Sessões Pequenas com e sem compressão

Tipo de Sessão	Precisão	Revocação	F1	Acurácia
Sem compressão	1.0000	0.7500	0.8333	0.7500
Com compressão	1.0000	0.7500	0.8333	0.7500

Assim como na tarefa anterior, a compressão não parece impactar a qualidade dos resultados do modelo para reuniões pequenas na detecção de posicionamentos, conforme ilustrado na Tabela 6.

Tabela 6: Média das Métricas de Desempenho para Detecção de Posicionamentos em Sessões Pequenas com e sem compressão

Tipo da Sessão	Precisão	Revocação	F1	Acurácia
Sem compressão	0.8125	0.8750	0.8333	0.8750
Com compressão	0.8125	0.8750	0.8333	0.8750

5 CONCLUSÕES E TRABALHOS FUTUROS

Este estudo contribuiu significativamente ao fornecer uma base de dados fundamental para esta e para outras análises futuras. O primeiro abrange todas as reuniões de comissões e sessões plenárias do Senado Federal do ano de 2023, além da construção manual de um *ground truth* por humanos, essencial para a avaliação precisa desses experimentos e outros similares.

Embora os resultados tenham revelado limitações em relação às reuniões de médio e longo porte, principalmente devido à sensibilidade da tarefa de atribuir posicionamentos individuais, reuniões de menor escala demonstraram resultados satisfatórios.

Além disso, pode-se destacar que a metodologia utilizada nesse trabalho, em particular na engenharia de *prompt*, foi construída para ser generalizada para outros contextos de debates, sejam assembleias de condomínio, assembleias universitárias, debates parlamentares, debates escolares, entre outros. Alterando os dados de entrasta e um parâmetro no *prompt*, é possível utilizar para esses contextos.

Este estudo busca estabelecer uma base sólida para investigações subsequentes. Assim, destacamos algumas direções de pesquisa que surgiram ao longo do desenvolvimento deste trabalho. São elas:

- **Avaliação de modelos mais recentes**: O GPT-3.5-Turbo já apresentou resultados otimistas, mas outros modelos mais recentes também podem ser avaliados nessas tarefas, como o própio GPT-4 [15] também da OpenAI, o recém lançado Llama 3 [16] da Meta, ou o Gemini [17] da Google.
- **Aumentar a quantidade de dados anotados para garantir uma melhor avaliação**: Seria benéfico aumentar a quantidade de dados anotados disponíveis para garantir uma avaliação mais robusta dos modelos. Isso poderia ser alcançado através da anotação de mais reuniões usando uma estratégia de amostragem estratificada para garantir que cada tipo de reunião esteja proporcionalmente representado, mantendo a equidade na avaliação.
- **Realizar *fine-tuning* em um LLM**: Um direcionamento que estudos futuros podem tomar é o *fine-tuning* de um LLM para realizar essas tarefas. Isso não foi feito neste estudo devido à quantidade de dados anotados necessários para o treinamento, que não está disponível atualmente.
- **Particionar os dados e adicionar como entrada para o modelo por partes**: Uma abordagem para lidar com a

[15] https://openai.com/index/gpt-4
[16] https://llama.meta.com/llama3/
[17] https://gemini.google.com/

sensibilidade do modelo em reuniões médias e longas pode ser dividir os dados em partes menores e enviar cada parte separadamente para análise. Isso pode ajudar a melhorar a precisão das classificações, reduzindo a carga de processamento do modelo.

- **Avaliar o impacto da compressão para reuniões médias e grandes**: Seria interessante investigar como diferentes níveis de compressão afetam a saída dos modelos para reuniões de diferentes tamanhos, fornecendo insights valiosos sobre como otimizar o desempenho do modelo para diferentes cenários.

6 AGRADECIMENTOS

O meu maior agradecimento vai para a minha mãe Elisabete e meu pai Ademirton, os meus maiores incentivadores nesse processo, que me ensinaram que eu era capaz de conquistar todos os meus objetivos e não mediram esforços para que eu pudesse ter uma educação de qualidade em toda minha vida e para que eu não desistisse do que eu almejava. Agradeço por terem se dedicado tanto por mim, pretendo honrar todo esse esforço. A minha prima Ismeralda, que sempre foi uma inspiração de disciplina e dedicação, me motivando a ser sempre uma versão melhor de mim. Agradeço também a toda minha família: avós, avôs, tias, tios, primos e amigas, que acreditaram no meu potencial desde a infância, e fizeram questão de fornecer todas as ferramentas necessárias para que eu pudesse alcançar meus sonhos através de uma educação de qualidade.

Quero agradecer também a toda a comunidade da Universidade Federal de Campina Grande (UFCG), que contribuiu imensuravelmente para minha formação acadêmica e como ser humano. Especialmente o meu orientador Cláudio Campelo, que me auxiliou não só nesse trabalho, mas na minha formação como um todo, fornecendo oportunidades sempre alinhadas com o que eu almejava, com uma excelente compreensão de como desenvolver melhor o potencial de todos os que coordena. Agradeço também ao professor Matheus Gaudencio, sempre com a sua percepção de ajudar os alunos em suas trajetórias, e que não me deixou desistir quando eu achava que não me encaixava no curso. Estendo meus agradecimentos a todos os professores, técnicos-administrativos e funcionários em geral responsáveis pelo curso de Ciência da Computação. Gostaria de agradecer também a toda equipe do Lacina, em especial ao meu *Txime*, Maria Eduarda, Alexandre Ribeiro, Filipe Lima, Matheus Lisboa, José Manoel, José Davi e Carlos Vinícius, com quem compartilhei um ano de pesquisa e desenvolvimento.

Não posso deixar de agradecer minhas amigas nesta jornada de graduação, em especial Andrielly Lucena, Anna Beatriz, Sheila Maria e Andressa Lucena. Com vocês, eu dividi intensos anos da minha vida e vocês estiveram ao meu lado, me dando todo suporte, carinho e torcendo por mim. Agradeço também a meus amigos Nícolas Moreira, Igor Farias, Mateus Matias, Arthur Alves, Henrique Lemos e Henrique Lopes, que tornaram os anos de graduação menos difíceis e mais divertidos.

Agradeço imensamente a Davi Sousa, meu namorado e companheiro na jornada desse trabalho, que às vezes acredita mais no meu potencial do que eu mesma e que sempre me incentiva a correr atrás dos meus objetivos, admiro você não só como futuro cientista da computação, mas como ser humano.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. *arXiv preprint arXiv:2004.03720* (2020).

[2] Helen Cavalcanti and Claudio Campelo. 2024. Dataset of Brazilian Federal Senate Session Transcriptions From 2023 with Relevant Topics and Stance Detection Annotations. https://doi.org/10.5281/ZENODO.11106904

[3] Matheus Alves dos Santos. 2024. Modelagem de tópicos na estimativa de pontos ideais baseados em discursos de parlamentares.

[4] Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. *arXiv preprint arXiv:2310.06839* (2023).

[5] Mircea-Luchian Pojoni, Lorik Dumani, and Ralf Schenkel. 2023. Argument-Mining from Podcasts Using ChatGPT. In *In procs. of the Workshops at International Conference on Case-Based Reasoning (ICCBR-WS 2023) co-located with the 31st International Conference on Case-Based Reasoning (ICCBR 2023), Aberdeen, Scotland, UK*, Vol. 3438. 129–144.

[6] Myrthe Reuver, Suzan Verberne, and Antske Fokkens. 2024. Investigating the Robustness of Modelling Decisions for Few-Shot Cross-Topic Stance Detection: A Preregistered Study. arXiv:cs.CL/2404.03987

[7] Patricia D Santos and Denise H Goya. 2021. Automatic twitter stance detection on politically controversial issues: A study on covid-19's cpi. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*. SBC, 524–535.

[8] İlker Gül, Rémi Lebret, and Karl Aberer. 2024. Stance Detection on Social Media with Fine-Tuned Large Language Models. arXiv:cs.CL/2404.12171