Do you know what your senator advocates for in the committees they participate in? An LLM-based approach to topic and stance detection in parliamentary discussions

Helen Bento Cavalcanti[1], **Claudio E. C. Campelo**[1]

[1]Systems and Computing Department
Federal University of Campina Grande (UFCG)
Campina Grande – PB – Brazil

helen.cavalcanti@ccc.ufcg.edu.br, campelo@dsc.ufcg.edu.br

Abstract. *The legislative power in Brazil faces challenges in making discussions more accessible to the population, which is essential for strengthening democracy. Although stenographic notes of the Senate and House of Representatives committee meetings are publicly available, their length and volume make it impractical for citizens to follow what actually happens in those meetings. Therefore, a tool that can automatically extract useful and summarized information from these discussions would be transformative, empowering voters to monitor their representatives more effectively. This study investigates the efficacy of Large Language Models (LLMs) for detecting relevant topics and stances of parliamentarians. We conducted experiments using GPT-3.5-Turbo to interpret shorthand notes from the Federal Senate in 2023. The results were promising, with an average accuracy of 70% and 60% for topic and stance detection, respectively.*

1. Introduction

Humans are inherently social beings, endowed with the ability to communicate, debate, and defend their viewpoints. Since ancient Greece, where agoras were venues for intense debates, numerous settings provide spaces for discussions and exchanges of ideas. In these situations, there is often an audience that, while not directly participating in the discussions, has a significant interest in what has been debated. This is because decisions made in these debates can have a direct or indirect impact on their lives. Examples include condominium assemblies, university assemblies, and parliamentary debates.

The Brazilian Legislature represents a peculiar and highly interesting scenario, as the population elects its representatives and seeks to follow their activities, positions, debates, expenditures, and other actions. Making these discussions more accessible to the public is essential for strengthening democracy and promoting social well-being, addressing the important and urgent challenges of contemporary society. In this context, computing emerges as a fundamental tool to assist citizens through techniques such as data mining, analysis, and visualization. Several initiatives already use innovative strategies in this regard. An example is the Vidinha de Balada project [1], which aims to evaluate

[1]https://www.jusbrasil.com.br/noticias/hackfest-cgu-incentiva-uso-da-tecnologia-para-combate-a-corrupcao-e-exercicio-da-cidadania/479985700

the expenditures of federal deputies and relate them to their performance in the Chamber. Another case is the Empenhados blog [2], which employs data analysis and visualization to provide relevant information to the public, such as data on public transportation in Paraíba and variations in product prices in public tenders.

However, monitoring committee debates and daily plenary sessions of the legislature is still challenging for the population, specially due to their length and volume. This gap gives greater freedom to representatives, as the difficulty in monitoring reduces societal demands. But this challenge is not exclusive to this legislative scenario; it can be generalized to other types of debate already mentioned.

In this context, this work aims to evaluate the potential of Large Language Models (LLMs) for tasks such as detecting relevant topics in debates and positioning of participants. For this purpose, an experimental scenario was delineated using debates from the Brazilian Senate. These debates generate stenographic notes [3], which comprise transcriptions of recorded speeches during plenary sessions and committee meetings.

LLMs are artificial intelligence models developed to understand and generate human language, trained extensively on vast textual datasets to absorb patterns, contexts, and nuances of natural language.

The experiments conducted in our study aim to answer the following research questions:

- **Q1** - What is the effectiveness of LLMs, specifically the GPT-3.5 Turbo model, in detecting relevant topics in the context of debates in the Brazilian Senate?
- **Q2** - What is the effectiveness of LLMs, specifically the GPT-3.5 Turbo model, in detecting positions in the context of debates in the Brazilian Senate?
- **Q3** - Does compressing the input data of the model impact the effectiveness of the results?

Additionally, this work adds three main contributions to computer science: (1) A generalizable prompt resulting from an intensive prompt engineering process, to be applied in tasks of detecting topics and positions in debate scenarios as a whole, where it is possible to parameterize and choose the context to be applied; (2) An evaluation of this approach, investigating the effectiveness of using LLMs for the tasks mentioned in the context of the 2023 Senate meetings; (3) And a relevant database [Cavalcanti and Campelo 2024] for future work, composed of stenographic notes from the 2023 Senate meetings, with manual annotations to enable the evaluation of model performance in the tasks of interest.

2. Related Work

The increase in research on large language models spans various fields, including news recommendation, where identifying whether articles share the same viewpoint is crucial. A recent study [Reuver et al. 2024] analyzed two tasks (*Same Side Stance* and *Pro/Con*) using architectures of Large Language Models (bi-encoding and cross-encoding) integrated with Natural Language Inference (NLI). The results indicated that recommendation

[2]https://analytics-ufcg.github.io/empenhados/

[3]https://www12.senado.leg.br/noticias/materias/2011/07/22/notas-taquigraficas-permitem-ao-cidadao-acesso-rapido-ao-trabalho-do-senado

systems benefit from robust models that consider a variety of topics, allowing efficient analysis of positioning differences between articles and offering new perspectives on current debates.

In the context of social media, a recent study [İlker Gül et al. 2024] explores the use of LLMs to detect stances in zero-shot learning scenarios[4] and few-shot learning[5]. The results show that LLMs, such as LLaMa-2 and Mistral-7B, outperform baseline models, even with their smaller sizes compared to ChatGPT. The study highlights the effectiveness of LLMs in detecting stances and discussed topics on social media, comparing their effectiveness in their respective tasks.

Additionally, LLMs are being applied in Argument-Mining, focusing on the automatic extraction and identification of argumentative structures in natural texts. A study [Pojoni et al. 2023] used GPT-4 to extract arguments from podcasts, an important platform for exchanging ideas and debates on various topics. This approach is relevant to this work as it involves using LLMs to analyze similar contexts of idea exchange and opinions.

In the context of political textual data analysis, Natural Language Processing techniques have been essential. A recent Brazilian study [dos Santos 2024] evaluates the use of BERTopic and Text-Based Ideal Point techniques for modeling latent topics and estimating ideal points, applying them to characterize speeches and positions of Brazilian parliamentarians in the 55th and 56th Legislatures (2015-2022). The results demonstrate the feasibility and potential of these techniques to support new political studies in Brazil.

In another study [Santos and Goya 2021], unsupervised stance extraction techniques, topic modeling, and automated labeling were employed to analyze retweeted posts about the Covid-19 Parliamentary Commission of Inquiry (CPI). Similarities were calculated among the most active users, and positioning was detected using dimensionality reduction, clustering, topic modeling with contextual embeddings, and automatic clustering labeling based on recurring terms. The study generated a small number of clusters (2 to 3), with label uniformity exceeding 98

This study differs from others by not only detecting topics but also identifying the debaters' stances on these topics, considering various political contexts. Additionally, it employs advanced prompt compression techniques before applying an LLM, distinguishing it from the other cited works.

3. Methodology

This section details the methodology used to evaluate the model's performance on the proposed tasks, including database construction, models employed, and experiments conducted.

3.1. Data Extraction and Pre-processing

The objective of this work is to analyze relevant topics and participants' stances in parliamentary meetings. The meetings were transcribed and converted into textual data.

[4]Approach where a model performs tasks for which it was not explicitly trained, without examples in the inference phase.

[5]Technique that allows the model to make predictions for new classes based on few examples.

To achieve this, a database was created covering the sessions of the Brazilian Federal Senate[6] in 2023, encompassing both plenary sessions and committee meetings. The database was centralized with information from the meetings during the period of interest, extracted from the official Senate website. Data scraping aimed to obtain Stenographic Notes, which are transcriptions of speeches made by parliamentarians or guests during sessions, as described by the Chamber of Deputies website[7].

Data extraction and pre-processing were implemented in Python, using the Pandas[8] and BeautifulSoup[9] libraries. The source code is available in a public repository on GitHub[10], and the complete database can be publicly downloaded [Cavalcanti and Campelo 2024].

The extraction process was carried out as follows: initially, all URLs of stenographic notes for the 2023 meetings were collected. Next, to extract the content, each URL was accessed to obtain the unique session identifier, speaker's name, party affiliation, and speech text. Table 1 presents the detailed structure of the dataset produced.

Tabela 1. Dataset Structure

Attribute	Description
id session	Unique event identifier
speaker name	Name of the person who delivered the speech
party	Political party
speech	Speech

During pre-processing, "INDEFINIDO"tag was used when the party was not specified. For speeches by the president, the "PRESIDENTE"tag was replaced with the president's name in question. Additionally, session with ID 25779 was removed due to the unavailability of notes on the Senate website.

3.2. Prompt Compression

Language models interpret text as a sequence of numbers, known as tokens. Byte Pair Encoding (BPE) [Bostrom and Durrett 2020] encoding technique transforms text into tokens. Providing text to the model, it converts it into tokens. Each model has a context window, which is the maximum amount of tokens accepted as input. In some cases, the context can be extensive for the model's window, leading to the need for prompt compression strategies, which help reduce costs and latency.

There are various approaches to prompt compression. This study utilizes smaller models, trained to identify non-essential tokens [Jiang et al. 2023], allowing the model to understand inputs after compression. GPT-3.5-Turbo has a context window of 16,385 tokens and does not accept larger inputs. As shown in Figure 1, only 41 of the 203 sessions contain up to 16,000 tokens, considering a small variation in the number of tokens after compression.

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

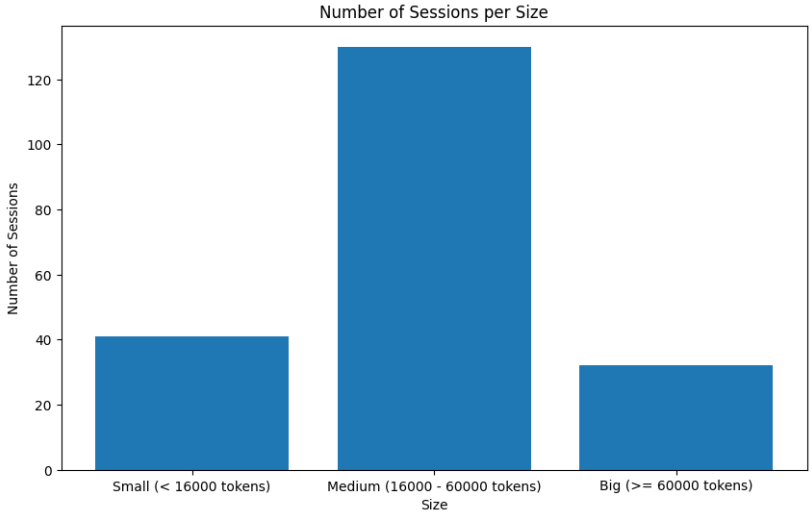

Figura 1. Distribution of session sizes

To overcome this challenge, a data compression strategy was adopted using LongLLMLingua [Jiang et al. 2023], a framework developed by Microsoft that facilitates rapid compression in extensive contexts. LongLLMLingua[11] is an innovative compression strategy that can achieve better performance across different tasks compared to the original prompt. Generally, compressed prompt can achieve better performance with less cost.

In implementing this strategy, essential parameters were defined. The main one is the target token, representing the maximum limit of tokens that the compression model should reach, set to 16,000 to optimize performance, as the effectiveness of compression decreases with increased compressed data. Two other significant parameters are context and question: context refers to additional context needed to answer questions based on meeting dialogues, while question concerns instructions provided to the language model, such as requests for information or specific questions. The final configuration of these parameters is detailed in Section 3.3.

3.3. Prompt Engineering

The proposed prompt was developed through a meticulous prompt engineering process, following the methodology proposed by Deeplearning.ai[12] in partnership with OpenAI. Several preliminary versions were evaluated. The final version of the prompt can be viewed in the GitHub repository [13].

The prompt is divided into three parts:

- It requests the model to behave as an expert in stance detection, explaining the concepts of being in favor, against, or neutral regarding a topic.
- A structured output in JSON format is used, specifying keys and values.
- The input text is explained with delimiters like “‘ (three backticks), indicating a variable called *context*.

[11]https://github.com/microsoft/LLMLingua
[12]https://www.deeplearning.ai/short-courses/chatgpt-prompt-engineering-for-developers/
[13]https://github.com/helenbc/tcc-notas-taquigraficas/blob/main/final$_prompt$

- A pipeline schema is structured with a sequence of action points to ensure the model performs all expected tasks, providing a list of individuals who made statements during the debate.

3.4. Experiment Definition

Two experiments were conducted. The first experiment was designed to address research questions **Q1** and **Q2**, which aim to measure the effectiveness of the GPT-3.5 Turbo model in the tasks of topic detection and stance detection in the context of debates at the Brazilian Senate. On the other hand, the second experiment aligns with **Q3**, seeking to understand if data compression impacts the effectiveness of the model's results.

The first experiment, as shown in Figure 2, is divided into two stages: text compression and inference using the GPT-3.5 Turbo model. In the first stage, each meeting is inserted as *context* into the compression model. The meeting's text is formatted with speeches arranged as follows: *"name": "speech"*, where the name is tagged with *llmlingua, compress=False* to prevent compression of this crucial segment for detecting the speaker's stance. The *question* parameter corresponds to the prompt described in Subsection 3.3, and the *target token* is set to 16,000, allowing for slight variation in the output token count.

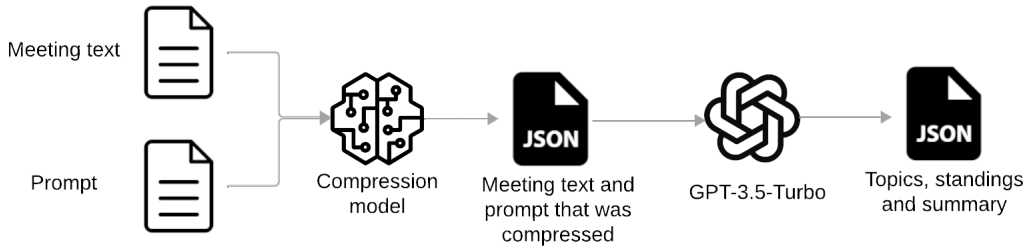

Figura 2. Diagram of steps for the first experiment

This process results in a JSON containing the compression of each meeting, which is then converted to a string and provided as input to the GPT model. The model returns, also in JSON format, the relevant topics discussed in the meeting and the stance of each participant regarding these topics.

The second experiment, as depicted in Figure 3, involves directly submitting smaller meetings (up to 15,000 tokens) to the GPT-3.5 Turbo model without compression, a number determined by the token space required for the prompt. The structure of the output remains the same as in the first experiment.

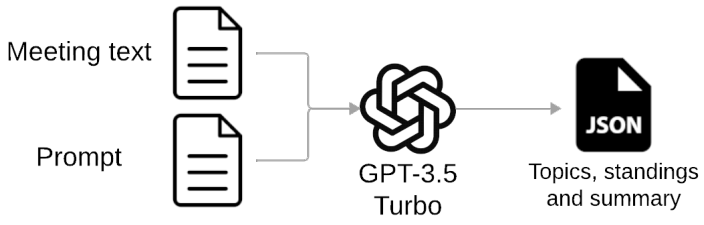

Figura 3. Diagram of steps for the second experiment

3.5. Model Evaluation

Model evaluation was divided into two distinct parts. In the first part, the focus was on the topics returned by the model, using metrics such as precision, recall, and F1-score for information retrieval. Precision represents the fraction of retrieved topics that are relevant, recall is the fraction of relevant topics that were retrieved, and F1-score is the harmonic mean of precision and recall. Classical terms like true positive, true negative, false positive, and false negative were used to calculate these metrics.

In the second part of the evaluation, which focuses on the stances of individuals regarding the discussed topics, metrics such as accuracy, precision, recall, and F1-score from the Scikit-learn library were used. Since the problem is multiclass, stances were mapped as **in favor** (0), **neutral** (1), and **against** (2). For each topic correctly predicted by the model, the stances of each parliamentarian were compared with manual annotations, and the mentioned metrics were calculated to evaluate the model's performance in stance detection.

4. Results and Discussions

This section presents and analyzes the results of the conducted experiments, addressing the research questions:

- **Q1** - What is the effectiveness of LLMs, specifically the GPT-3.5 Turbo model, in detecting relevant topics in the context of debates at the Brazilian Senate?
- **Q2** - What is the effectiveness of LLMs, specifically the GPT-3.5 Turbo model, in detecting stances in the context of debates at the Brazilian Senate?
- **Q3** - Does compression of input data impact the effectiveness of the results?

The experiments were conducted on a machine equipped with an RTX 2080 Ti GPU, which has 11 GB of GPU memory, crucial for the first experiment involving data compression. The chosen model was the quantized Llama-2-7b-Chat-GPTQ from the LLMLingua library, requiring a minimum of 8 GB of GPU memory.

Metrics were calculated using a dataset comprising six meetings, distributed into two groups of each size (small, medium, and large), totaling 1013 minutes. The meetings were manually annotated in a JSON format structurally similar to the model's output. Ideal stratified sampling could not be performed due to limitations in human resources for this task.

4.1. Q1 - What is the effectiveness of LLMs, specifically the GPT-3.5 Turbo model, in detecting relevant topics in the context of debates at the Brazilian Senate?

The first experiment revealed that the model demonstrates superior performance in detecting topics in smaller meetings, as indicated in Table 2. In larger meetings, there was a significant decrease in metrics, with precision reaching 100% for small meetings but recall dropping below 50% for medium and large meetings. This suggests that the model tends to correctly identify mentioned topics, especially at the beginning and end of meetings, but may require adjustments to improve performance in larger sessions.

Tabela 2. Average Performance Metrics for Topic Detection by Session Type

Session Type	Precision	Recall	F1-score	Accuracy
Small	**1.0000**	**0.7500**	**0.8333**	**0.7500**
Medium	0.7500	0.4804	0.5130	0.3471
Large	0.3750	0.0551	0.0959	0.0505
Overall Average	**0.7083**	**0.4285**	**0.4807**	**0.3825**

4.2. Q2 - What is the effectiveness of LLMs, specifically the GPT-3.5 Turbo model, in detecting stances in the context of debates at the Brazilian Senate?

In the first part of the first experiment, the model exhibited superior performance in detecting topics in smaller meetings, as shown in Table 2. As the size of the meetings increased, there was a significant drop in metrics, especially in larger meetings where recall was notably low.

In the second part of the experiment, focusing on stance detection, a similar pattern of superior performance in smaller meetings was observed, with accuracy nearing 90%. However, in medium-sized meetings, the performance was lower, closer to what was observed in large meetings.

Qualitative analysis revealed that especially in larger meetings, the model tended to categorize all participants with a single stance, even those who did not speak on the topic. This suggests a limitation in the model's ability to discern individual stances in more complex contexts.

Tabela 3. Example comparing real data with model output where most stances are neutral

Real Data	Model Output
"Bill No. 2.788, 2019": "Mrs. DAMARES ALVES": "FOR" "Mr. JORGE KAJURU": "FOR" "Mrs. ZENAIDE MAIA": "FOR"	"Bill No. 2.788, 2019": "Mrs. DAMARES ALVES": "NEUTRAL" "Mr. PRESIDENT RODRIGO PACHECO": "NEUTRAL" "Mr. EDUARDO GIRÃO": "NEUTRAL" "Mr. FLÁVIO ARNS": "NEUTRAL" "Mr. CONFÚCIO MOURA": "NEUTRAL" ... "Mr. MARCELO CASTRO": "NEUTRAL"

In the second example, related to the "House Substitute for Bill No. 4.727, 2020,"again, the real data indicates a stance in favor for certain members, but the model labels them as against, as evidenced in Figure 4. This represents a more significant error compared to the previous one because in addition to mislabeling those who did not speak on the topic, the model did the same for those who clearly spoke in favor of that topic.

Tabela 4. Example comparing real data with model output where most stances are against

Real Data	Model Output
"House Substitute for [...] No. 4.727, 2020": "Mr. PRESIDENT RODRIGO PACHECO": "FOR" "Mrs. SORAYA THRONICKE": "FOR"	"House Substitute for [...] No. 4.727, 2020": "Mrs. DAMARES ALVES": "AGAINST" "Mr. PRESIDENT RODRIGO PACHECO": "AGAINST" "Mr. EDUARDO GIRÃO": "AGAINST" "Mr. FLÁVIO ARNS": "AGAINST" "Mr. CONFÚCIO MOURA": "AGAINST" ... "Mr. MARCELO CASTRO": "AGAINST"

Tabela 5. Average Performance Metrics by Session Type

Session Type	Precision	Recall	F1-score	Accuracy
Small	**0.8125**	**0.8750**	**0.8333**	**0.8750**
Medium	0.5030	0.5918	0.5251	0.5918
Large	0.4640	0.4464	0.4541	0.4464
Overall Average	**0.5932**	**0.6377**	**0.6042**	**0.6377**

4.3. Q3 - Does data compression impact the effectiveness of model results?

Analyzing the second experiment directly linked to the third research question, it was found that data compression had no impact on the model's performance in detecting topics in small sessions, as shown in Table 6. Ideally, conducting additional tests with more extensive datasets and sessions of varying sizes, applying different levels of compression, would provide a more comprehensive and reliable evaluation to confidently assert that compression does not affect results.

Similarly to the previous task, compression does not appear to impact the quality of model results for small sessions in stance detection.

Tabela 6. Average Performance Metrics for Topic and Stance Detection in Small Sessions with and without Compression

Task	Session Type	Precision	Recall	F1-score	Accuracy
Topic Detection	No Compression	1.0000	0.7500	0.8333	0.7500
	With Compression	1.0000	0.7500	0.8333	0.7500
Stance Detection	No Compression	0.8125	0.8750	0.8333	0.8750
	With Compression	0.8125	0.8750	0.8333	0.8750

This study provided a foundational dataset by encompassing all committee meetings and plenary sessions of the Federal Senate in 2023, with manually constructed ground truth by humans, essential for accurate evaluations of future experiments.

While there were limitations in assigning individual stances in medium and large meetings, smaller-scale meetings showed satisfactory results. The methodology developed in prompt engineering can be generalized to various debate contexts, such as condominium, university, parliamentary, and school assemblies, allowing adaptations through modification of input data and prompt parameters.

This study aims to establish a solid foundation for future research, highlighting the following emerging research directions:

- **Evaluation of newer models**: Evaluate models like OpenAI's GPT-4 [14], Meta's Llama 3 [15], and Google's Gemini [16] to enhance performance in similar tasks.
- **Increase in annotated data quantity**: Increase the amount of annotated data using stratified sampling to ensure proportional representation of each type of meeting, improving model assessment.
- **Fine-tuning in an LLM**: Explore fine-tuning of language models to adapt them to the specific tasks of this study, enabling better performance in various contexts.
- **Partitioning data for segmental analysis**: Divide data into smaller segments and analyze them separately to potentially increase accuracy of classifications in medium and large meetings, while reducing model processing load.
- **Evaluation of compression impact in medium and large meetings**: Investigate how different levels of data compression affect model outputs to optimize performance in different meeting scenarios.

Referências

Bostrom, K. and Durrett, G. (2020). Byte pair encoding is suboptimal for language model pretraining. *arXiv preprint arXiv:2004.03720*.

Cavalcanti, H. and Campelo, C. (2024). Dataset of brazilian federal senate session transcriptions from 2023 with relevant topics and stance detection annotations.

dos Santos, M. A. (2024). Modelagem de tópicos na estimativa de pontos ideais baseados em discursos de parlamentares.

Jiang, H., Wu, Q., Luo, X., Li, D., Lin, C.-Y., Yang, Y., and Qiu, L. (2023). Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. *arXiv preprint arXiv:2310.06839*.

Pojoni, M.-L., Dumani, L., and Schenkel, R. (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*, volume 3438, pages 129–144.

Reuver, M., Verberne, S., and Fokkens, A. (2024). Investigating the robustness of modelling decisions for few-shot cross-topic stance detection: A preregistered study.

Santos, P. D. and Goya, D. H. (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*, pages 524–535. SBC.

İlker Gül, Lebret, R., and Aberer, K. (2024). Stance detection on social media with fine-tuned large language models.

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