GoViz: A Visualization Tool for Empowering Transparency in Government Speech.

Larissa Guder[a*], *João Paulo* Aires[a], *Isabel H.* Manssour[a], *Dalvan* Griebler[a]

[a]Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, Brazil, larissa.guder@edu.pucrs.br

Submitted: 31 January 2025, Revised: 26 March 2025, Accepted: 21 April 2025, Published: 19 May 2025

Abstract. Public speech from government figures often describes relevant actions that can impact the population's lives. However, most people do not have time and access to analyze and understand public speech. Such a scenario narrows the participation of the people in the main discussions, which leads to multiple misunderstandings. In this work, we propose GoViz, a tool that automatically produces visual representations to outline governmental speeches regarding the subject, its main actors, and how they connect to the discussion topics. GoViz processes natural language from speech transcriptions in a pipeline that identifies part-of-speech elements, named-entities, and the relation between persons, making speech content more accessible and insightful. Using publicly available data, we evaluate our tool in two different languages (Portuguese and English). The results demonstrate that the visualizations from both data facilitate understanding the speech content. Thus, our main contribution is to encourage the participation of citizens in parliamentary issues, allowing a simplified and visually engaging avenue to access long speeches and fostering improved communication between parliamentarians and the population.

Keywords. natural language processing, data visualization, digital government
Research paper, DOI: https://doi.org/10.59490/dgo.2025.954

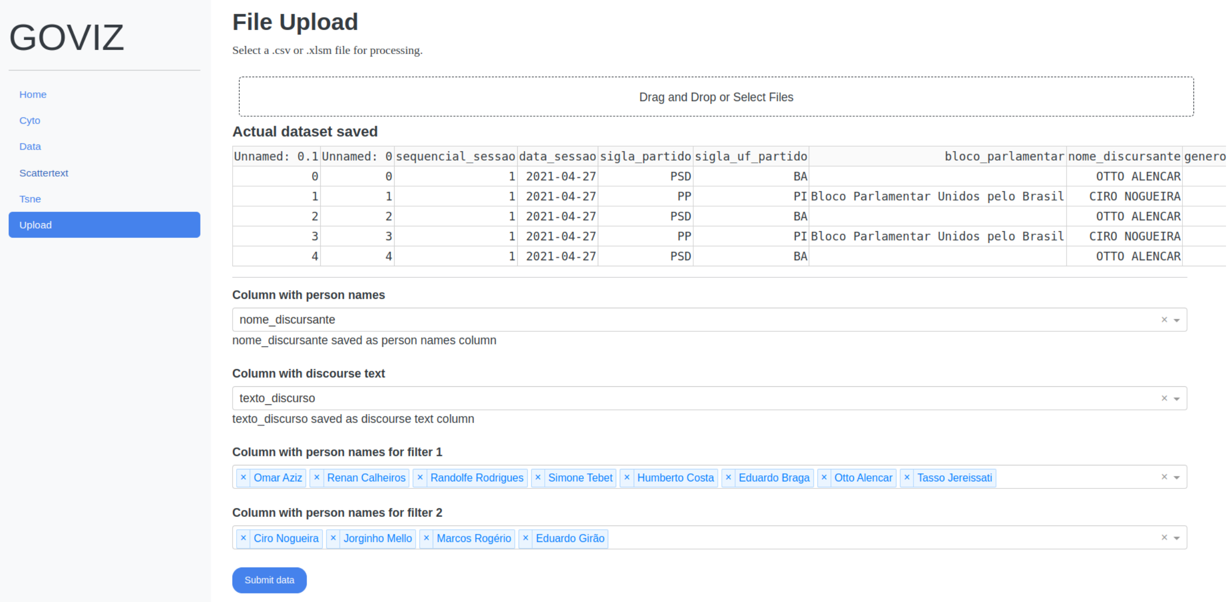
1. Introduction

Most representative democracies define and discuss laws and public measures in parliaments. There, politicians elected by the people vote and argue in favor of the interests of their voters. Although these politicians are the ones to accomplish and propose discussions, the public must also follow and question the decisions and which topics must be covered. However, the language used in these discussions and the decisions made are often hard to comprehend, making them inaccessible to the population. In the analysis of Bernardes and Bandeira, 2016 about parliamentary websites in Brazil and the United Kingdom, one of the main problems found is that even though there is a lot of information available, it can be an exhaustive task to explore and understand such information. Thus, although there is information about these discussions, they still cannot reach the population in a way that facilitates understanding.

A simple way to facilitate the understanding of parliament discussions is through visualizations. Generating images and interactive charts showing the main topics in a discussion can make it easier to follow them. Existing approaches (Cantador et al., 2021; Palma et al., 2021) propose generating visualizations from debates that try to explain most of the discussion. However, they often need extra information from someone who already knows about the subject to guide the generation of these visualizations. Considering that most people do not know about the topic, one of the main features of a visualization tool should be the capacity to process information without any previous knowledge about it.

The Global Parliamentary Report (Inter-Parliamentary Union, 2022) defines the parliamentary engagement in five different points: (1) information, (2) education, (3) communication, (4) consultation, and (5) participation. The citizens' engagement starts with understanding what is happening in the government. In this work, we introduce GoViz, a visualization tool that automatically processes discourses to provide insightful visualizations that facilitate their analysis. Our main objective with GoViz is to provide a practical application of natural language processing combined with visualization techniques to engage the citizens at the information level. Alternatives like this are necessary because more than making information publicly available is needed. It is necessary to grant that the information is relevant, reliable, timely, and comprehensive, and mainly that it can be understood by everyone in society, regardless of their background, status, or abilities (Inter-Parliamentary Union, 2022).

GoViz uses natural language processing to detect different patterns in speech transcriptions. It can receive structured documents as input and automatically identify columns containing speech and those containing people's names. By connecting people's names with their discourses, GoViz can reveal the main subjects discussed by every speaker and how frequently they talk about other people. Currently, GoViz can work with both English and Portuguese languages, which allows the processing of different types of texts. We evaluate GoViz using two publicly available datasets, one in English and the other in Portuguese. They bring political discussions about sensitive topics that are of people's interest. We perform a series of analyses of the generated visualizations and describe how they connect to the main subject of the datasets and their main limitations.



Select a .csv or .xlsm file for processing.

Actual dataset saved

Unnamed: 0.1	Unnamed: 0	sequencial_sessao	data_sessao	sigla_partido	sigla_uf_partido	bloco_parlamentar	nome_discursante	genero
0	0	1	2021-04-27	PSD	BA		OTTO ALENCAR	
1	1	1	2021-04-27	PP	PI	Bloco Parlamentar Unidos pelo Brasil	CIRO NOGUEIRA	
2	2	1	2021-04-27	PSD	BA		OTTO ALENCAR	
3	3	1	2021-04-27	PP	PI	Bloco Parlamentar Unidos pelo Brasil	CIRO NOGUEIRA	
4	4	1	2021-04-27	PSD	BA		OTTO ALENCAR	

Column with person names

nome_discursante

nome_discursante saved as person names column

Column with discourse text

texto_discurso

texto_discurso saved as discourse text column

Column with person names for filter 1

× Omar Aziz × Renan Calheiros × Randolfe Rodrigues × Simone Tebet × Humberto Costa × Eduardo Braga × Otto Alencar × Tasso Jereissati

Column with person names for filter 2

× Ciro Nogueira × Jorginho Mello × Marcos Rogério × Eduardo Girão

Fig. 1 – Upload screen.

GoViz's contribution is two-fold: first, it increases the direct participation of citizens in parliamentary issues through access to long speeches in a simplified and visual way; and second, GoViz implementation code is publicly available on https://github.com/NLP-PUCRS/GoViz, which allows other people to modify it to work in different contexts and languages. This paper is organized as follows: Section 2 introduces the main concepts concerning this work, such as natural language processing and visualization techniques. We also present some related work that shares some aspects with our approach. Section 3 describes the methods used for building GoViz. Here, we explain how we build GoViz, detailing its different models. In Section 4, we describe the main experiments and the results obtained. We introduce the scenarios and how we extract information from the selected datasets. Given our results, Section 5 discusses the main findings of this work and its limitations. Finally, in Section 6, we summarize our work and bring some perspectives on future work.

2. Background

GoViz is a tool built on two main concepts: Natural Language Processing (NLP) and Data Visualization. From NLP, we leverage the main techniques for extracting textual information. On the other hand, from data visualization, we use existing approaches for summarizing information into visual representations. In this section, we detail the visualization concepts and present related work.

2.1. Text Visualization

When performing data analysis, extracting information from textual data can be difficult. As text can carry relevant information, it is necessary to identify where this information is and how we can summarize it. Text visualization techniques can help with this task. We can consider visualization as a method for communicating information through an image or interactive chart (Ward et al., 2015). Such an image or chart can carry as much information as possible, making it easier for a human to understand the context it is describing. Text visualizations have corpora as a basis and can focus on words, sentences, paragraphs, or entire documents.

One of the first steps to process text is converting it into vector space. In the vector space, we have information about the presence and frequency of words in sentences/documents. There are many ways to obtain a vector that represents a text. Counting the frequency of terms (bag-of-words) in a text is one of the most basic ways to get a vector. Each position in the vector will represent a word with a number indicating the frequency of such a word. With this initial representation, we can already create visualizations that show frequent terms in a text. Other representations can involve the frequency of words in a document weighted by their presence in other documents in a corpus, such as Term-Frequency inverse Document-Frequency (Rajaraman & Ullman, 2011).

A common way to visualize a document is by displaying the frequency of words. Word cloud is a visualization technique that shows the word frequency of a document. It highlights more frequent words by increasing its font size and color darkness (Ward et al., 2015). The result is a cloud-shaped structure containing such words. Similarly, Word Tree is a technique that shows the frequency of words and the context in which they are inserted and allows easy exploration (Wattenberg & Viégas, 2008). Thus, given a document and a target word, it provides a visualization where the target word is the root and the sentences where it appears are the branches.

These techniques demonstrate how we can organize textual information to understand the context without reading it. In addition to Word Cloud and Word Tree, other text visualization techniques can support the representation of connections in the text, such as the arc diagram (Heer et al., 2010), or analyze feature values across the text, such as the literature fingerprinting presented by (Keim & Oelke, 2007).

An alternative to visualize text representations and other high-dimensional elements is using t-SNE (Van der Maaten & Hinton, 2008). This technique allows the visualization of multidimensional points in two or three dimensions. When converting from a multidimensional to a two-dimensional representation, t-SNE can preserve the approximate distance between similar points and non-similar ones. Modifying the concept of Stochastic Neighbor Embedding (SNE) (Hinton & Roweis, 2002), t-SNE proposes a new method to identify the similarity of low-dimensional points using the Student-t distribution. Besides, they propose a new cost function to approximate the real distance between points of low-dimensional representations converted from high-dimensional ones. In this work, we use text visualizations to facilitate people's understanding of speeches from politicians.

2.2. Related Work

Transparency, participation, and collaboration can be considered the principles of the open government (o-government) (Harrison et al., 2011; Quintero-Angulo et al., 2020). Transparency is crucial in making it possible for citizens to participate and collaborate with the government. However, making large datasets publicly available is not enough. This data must be accessible, clear, and easy to understand for citizens. Government data can have different formats and sources. It can be financial values, statistical, and textual, such as speech transcriptions. Manual summarization and information retrieval from textual data are laborious tasks since it is necessary to read everything to understand the subject of the data. Thus, NLP becomes a valuable alternative to extract information from texts automatically.

There are several research initiatives to facilitate the comprehension of a textual dataset, such as chatbots that extract only the desired information (Cantador et al., 2021) and the application of topic mapping from text (Palma et al., 2021). However, these approaches still require prior knowledge of the subject to extract information correctly.

When we refer to speech transcriptions, specifically of debates and public discussions, we can use data visualization techniques to facilitate the understanding of their subjects or to extract key information from these discussions. Khartabil et al., 2021 propose a tool for argument visualization that allows us to determine which points each person is defending. The authors proposed three novel visualizations: Stacked Boxes, Sunburst Pop-Up, and F+C Icicle. Their approach is suitable for deep analyses but requires the user to pass some arguments. Our approach focuses on providing a lean summarization of the content from speech transcriptions.

When analyzing speeches from politicians, it is important to see the agreements and disagreements in political discussions (Volkovskii & Filatova, 2022) and polarization (Diaz et al., 2022). With this information, citizens can use their critical sense and build their vision about the subject in discussion. To facilitate the visualization of different points of view, graphs such as the scatter-text proposed by Kessler, 2017 allow us to compare two groups of people. The scatter-text allows us to clearly see the polarization between two groups and which terms they commonly address. Because of this, we have incorporated this type of graph into our tool. Its use is available when processing documents in Portuguese and English.

Focusing on open government data, Steinbauer et al., 2016 and Bönisch et al., 2023 propose visualizations to explore transcribed parliamentary speeches. Steinbauer et al., 2016 explores Austrian parliament data, proposing a tool to visualize the relationships between politician discourse and parliamentary groups. The authors present the politician's profile and overview through visualizations, including the politician's speech, club, absence, and a politician relation graph in the parliament. While Steinbauer et al., 2016 uses the sentiment expressed in the speech towards topics discussed, we analyze the content of the speech to define the relations. Bönisch et al., 2023 proposes a web application to explore the minutes of plenary proceedings, agenda items, and polls of the German government data. The proposed tool contains a topic analysis, focusing on a specific speaker, party, or fraction on a specific topic. The Deputy Inspector analyzes a speech in conjunction with complementary data from background information, remarks, and poll results. In contrast to Bönisch et al., 2023 tool, which is specific to German government data, our tool is generalist, accepting various sources of data in Portuguese or English.

3. GoViz Description

In this section, we describe the methodology used to develop GoViz. We explain the data input supported, the main techniques used to process the text, and the available visualizations to analyze it. Figure 2 illustrates all the processing stages described in the following sections.

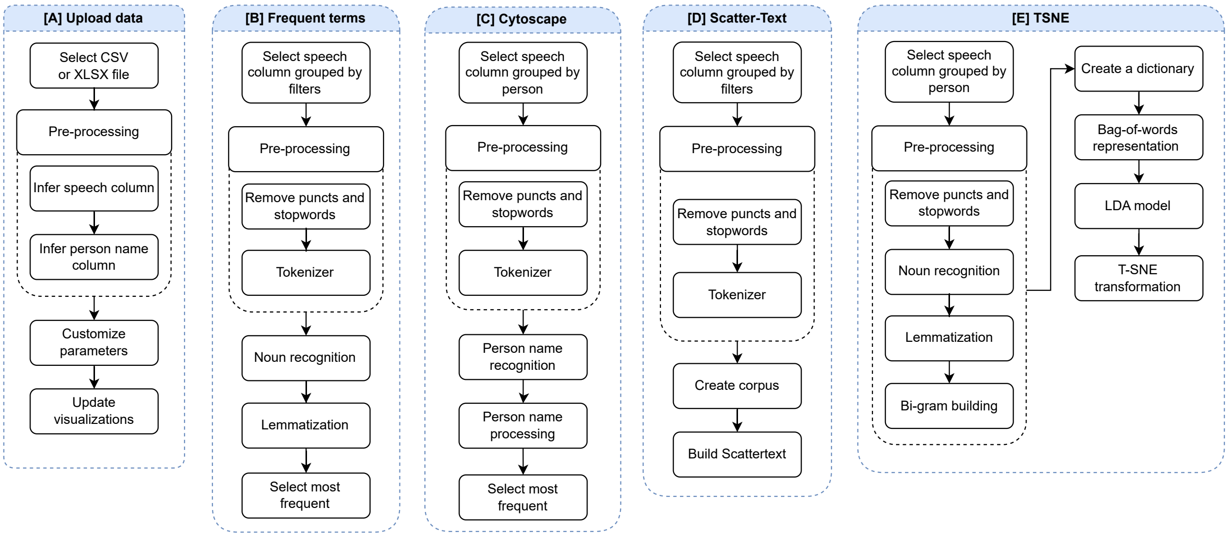


Fig. 2 – Processing stages for chart generation.

3.1. Methodology

For this work, we initially analyzed related work involving the visualization of governmental data. Then, we studied NLP methods and visualization techniques to specify the requirements and functionalities of GoViz.

With this initial study, we defined the type of information we aim to extract from governmental speeches and which NLP models we would employ to obtain them. Finally, given the textual information from the speeches, we selected the best visualization techniques that allowed us to summarize the data.

We process textual data using SpaCy (Montani et al., 2020) and Gensim (Rehurek & Sojka, 2011), while for manipulating the dataset, we use Pandas[1]. For the visualizations, we use the Plotly library[2] to plot the graphs and Dash[3] to build the tool. For this work, we use Spacy for most natural language processing tasks. We selected this library due to its ability to process different languages accurately. Since we want GoViz to work on both Portuguese and English properly, Spacy became a reliable option. Spacy provides the obtained accuracies for each task in different languages. We summarize the results in Table 1. As we can see, both languages have an accuracy higher than 95% for both Tokenization and Part-of-Speech Tagging, while having an F-score higher than 85%. We can see the full results for English[4] and Portuguese[5] in their page. As source for training the models, Spacy uses OntoNotes 5[6], ClearNLP Constituent-to-Dependency Conversion[7], WordNet 3.0[8], and Explosion Vectors (OSCAR 21.09 + Wikipedia + OpenSubtitles + WMT News Crawl)[9]. While for Portuguese, they use UD Portuguese Bosque v2.8[10], WikiNER[11], and Explosion fastText Vectors (cbow, OSCAR Common Crawl + Wikipedia).

Language	Task	Result
English	Tokenization	1.0
	Part-of-Speech	0.97
	Named Entity Recognition	0.85
Portuguese	Tokenization	1.0
	Part-of-Speech	0.97
	Named Entity Recognition	0.89

Tab. 1 – Spacy results for NLP tasks in both English and Portuguese. For both Tokenization and Part-of-Speech Tagging, they provide an accuracy result, while for Named Entity Recognition, they use F-score.

3.2. Input data

The main idea behind the proposed approach is that the user can use different textual datasets to generate visualizations to extract information without the need to read the entire transcription. Textual data can be collected from various sources and stored in different ways. Providing a standard input format is essential to allow a correct visual representation. Thus, GoViz accepts structured data in `csv` and `xlsx` formats. The dataset must contain a column with the person's name and one with textual data. Since our goal is to generate visualizations from speeches, it is important to have information about who is saying the transcribed text. Regarding the textual language, it can be either in Portuguese or English.

To facilitate the use of GoViz by its users, we built an interface, illustrated in Figure 1. It contains an upload function that receives a `csv` or `xlsx` file as input. Once the file is uploaded, we process the textual data using SpaCy and Gensim. Different processing is done for each visualization, as described in Figure 2. We store the processed data in a JSON file that GoViz reads to generate each chart every time the application starts.

3.3. Processing stages

GoViz has five main processing stages. The first one processes the uploaded data. Initially, GoViz pre-processes the uploaded file to infer columns that contain the name of the speaker and the one containing the speech. However, the user can change the automatically selected columns if the inference fails to find them correctly. To

[1] https://pandas.pydata.org/
[2] https://plotly.com/
[3] https://dash.plotly.com/
[4] https://spacy.io/models/en#en_core_web_md-accuracy
[5] https://spacy.io/models/pt#pt_core_news_md-accuracy
[6] https://catalog.ldc.upenn.edu/LDC2013T19
[7] https://github.com/clir/clearnlp-guidelines/blob/master/md/components/dependency_conversion.md
[8] https://wordnet.princeton.edu/
[9] https://github.com/explosion/spacy-vectors-builder
[10] https://github.com/UniversalDependencies/UD_Portuguese-Bosque
[11] https://figshare.com/articles/dataset/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500

make the inference, we first convert the dataset into a Pandas DataFrame and verify the column type, removing everything with a data type other than text. From the remaining columns, we select the one containing the largest textual length as the speech column. Once selected, we use SpaCy to perform named-entity recognition on the other columns to identify the person's name column. We select the first word from each row in a candidate column and check if it is a named-entity object identified as a person. The column with the highest occurrence of person names is then selected. Before submitting the data for the next stages, the user can define two groups of person names to compare their speeches. Figure 2[A] illustrates the steps from the first processing stage.

A simple way to identify patterns in textual data is by obtaining the frequency of words. We select data from the provided file based on the groups defined in the data upload stage to extract the frequency of terms. If the user did not provide any groups, we consider the top terms from the automatically identified column containing textual data. When defined by the user, we split the text based on the two informed groups. First, we remove stop words and punctuations to pre-process and convert the raw text into tokens to feed the Spacy model that performs part-of-speech tagging. We only consider nouns for counting the term frequency, as they contain more relevant information, such as things, places, and persons. After filtering the terms, we convert them to lemmas and select only the top 20 most frequent nouns. We illustrate this processing stage in Figure 2[B].

We process data for the third processing stage to obtain the person's names mentioned in each speech. First, we group each speech by the speaker. Then, we pre-process the data by removing punctuation and stop-words and convert it into tokens. We feed the named-entity recognizer Spacy model with the cleaned data. From the output, we keep only the tokens recognized as Person entities. Finally, we get the people's names cited by their frequency. Figure 2[C] shows the steps of this processing stage.

To generate the Scatter-text visualization (see Section 3.4), in the fourth processing stage, we need to create a corpus of frequent terms for two separate groups. This process only occurs when the user previously indicates two groups for comparison. The idea is to visually compare the main terms that each group brings into the discussion. Thus, given each group's speeches, we first use the same pre-processing steps described in the previous stages. Then, using the library provided by Kessler, 2017, we create a corpus containing tokens that we will use to create groups of words and the most frequent terms mentioned by the participants of each group. Figure 2[D] shows the steps to perform this processing stage.

Finally, the fifth processing stage manipulates textual data to serve as input to the t-SNE model (see Section 3.4). In this process, the first step is to pre-process speeches by selecting them by their speaker. Then, we remove punctuation and stop words. To obtain the part-of-speech of the terms, we use Spacy. Once we have the part-of-speech, we select only the nouns, convert them into lemmas, and create bi-grams out of them. With the cleaned data, we feed a Latent Dirichlet allocation (LDA) model provided by Gensim. This model tries to identify the topics given a set of documents. For our specific case, we want to identify topics given in the speeches in the dataset. The number of topics to be created is customizable in the script. With the output of LDA, we made the transformations for t-SNE. We illustrate these steps in Figure 2[E].

3.4. Visualizations

GoViz provides four interactive visualizations: treemap, network, scatter text, and t-SNE. Each visualization has a different level of user interaction that facilitates a more detailed analysis of the numbers and terms present in the visualization. To use all GoViz resources, the user must specify two groups of people's names with the aim of comparing the speeches. If not informed, GoViz still brings the treemap and t-SNE visualizations.

Focusing on understanding the main terms mentioned by speakers, we use a treemap to show the top 20 nouns. If the user specifies the two groups of people's names, GoViz shows the top 20 for each group, using two different colors, as exemplified in Figure 3. Otherwise, it shows the general top-20, as presented in Figure 7. The size of each rectangle represents the frequency of the term. For this specific figure, we omitted words that do not fit the rectangle, as we increased the font size for better visualization.

The treemap allows us to overview the main terms, but it is limited to 20 of them. To provide a deeper analysis, we use the scatter text visualization. To build the scatter text, the user must specify two groups of persons that will have their speeches compared, considering the terms they use. Thus, the visualization shows terms

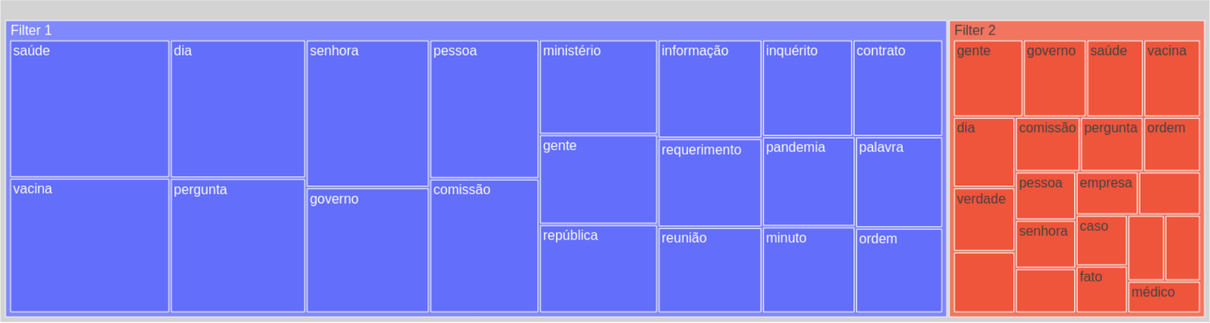
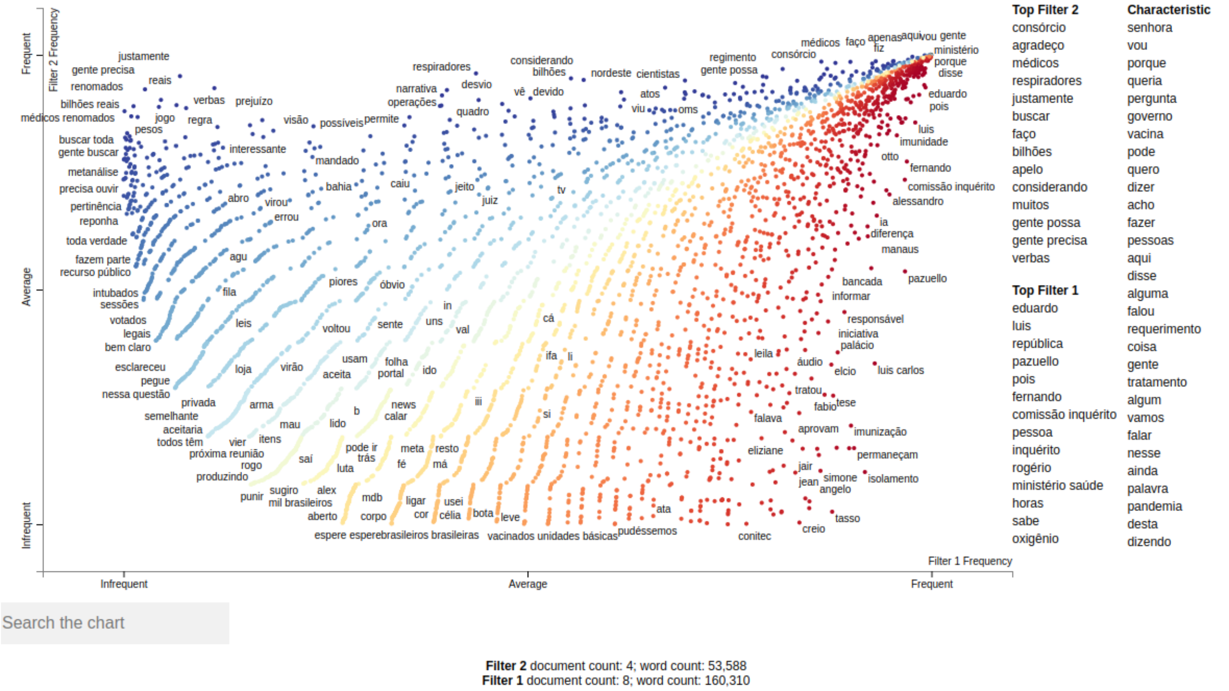


Fig. 3 – Most frequent nouns in the PCI Covid dataset.

used by one group, by the other, and those used by both. As presented in Figure 4, the terms in the center are used by both sides. On the top-left, we have the most used terms by group 2, while on the bottom-right, we have the most used terms by group 1. This is important to show where the discourse can be similar and at which points they diverge.



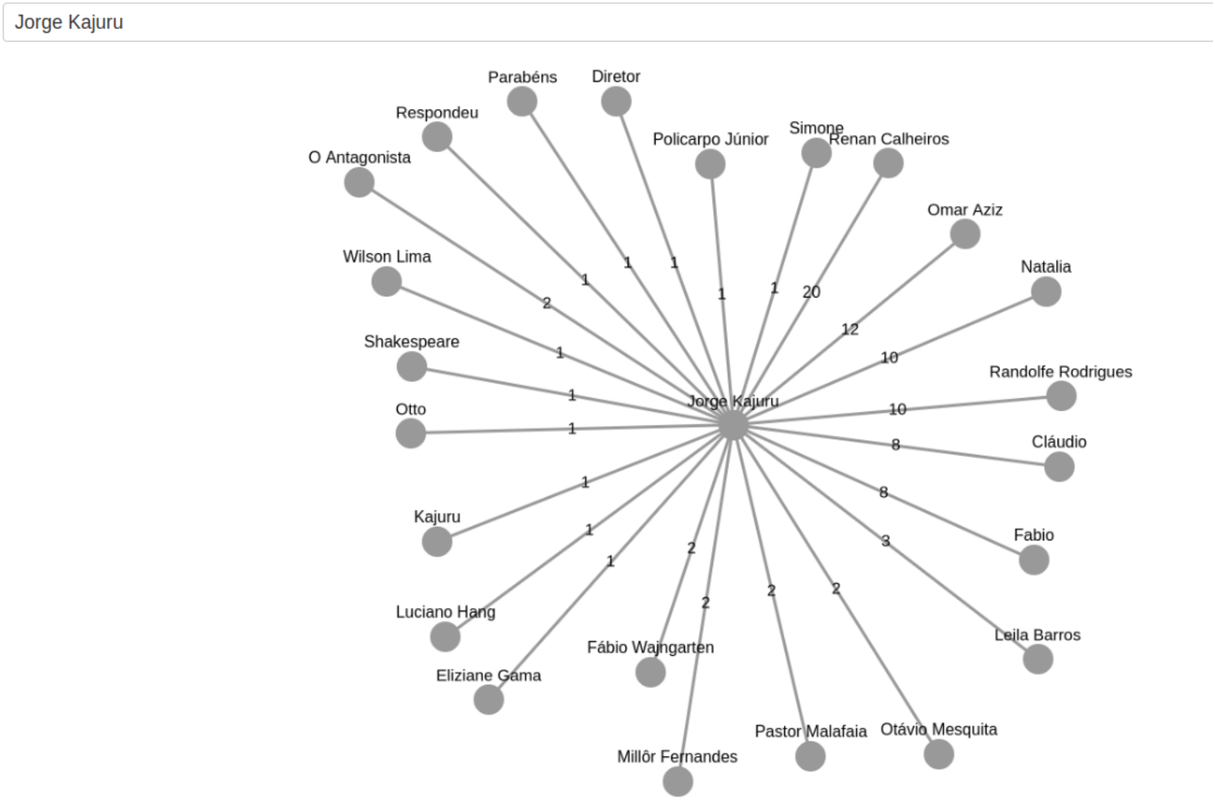
Top Filter 2	Characteristic
consórcio	senhora
agradeço	vou
médicos	porque
respiradores	queria
justamente	pergunta
buscar	governo
faço	vacina
bilhões	pode
apelo	quero
considerando	dizer
muitos	acho
gente possa	fazer
gente precisa	pessoas
verbas	aqui
	disse
Top Filter 1	alguma
eduardo	falou
luis	requerimento
república	coisa
pazuello	gente
pois	tratamento
fernando	algum
comissão inquérito	vamos
pessoa	falar
inquérito	nesse
rogério	ainda
ministério saúde	palavra
horas	pandemia
sabe	desta
oxigênio	dizendo

Filter 2 document count: 4; word count: 53,588
Filter 1 document count: 8; word count: 160,310

Fig. 4 – Scatter-text visualization with PCI Covid dataset.

To verify which people were cited during the speeches, we used Cytoscape, a network visualization provided by Dash. It shows the relationship between the speaker and the people mentioned. Figure 5 shows an example with the speaker's name in the center and the connections showing how many times he/she mentioned each person.

Finally, the visualization generated by t-SNE (Figure 6) aims to show the relationship between topics and who is talking about them. Each circle represents a speaker, and the size of each circle means how many times that speaker talked about the topic. It is possible to see the speaker's name with a mouse click on the circle or selection. Each color represents a different topic, and their position in the graph means how close each group is to the other. Using t-SNE, we can see a complement from the treemap, where we can focus on each speaker and understand which points they are making during their speech.



Jorge Kajuru

Fig. 5 – Cytoscape visualization with PCI Covid dataset.

4. Case Studies

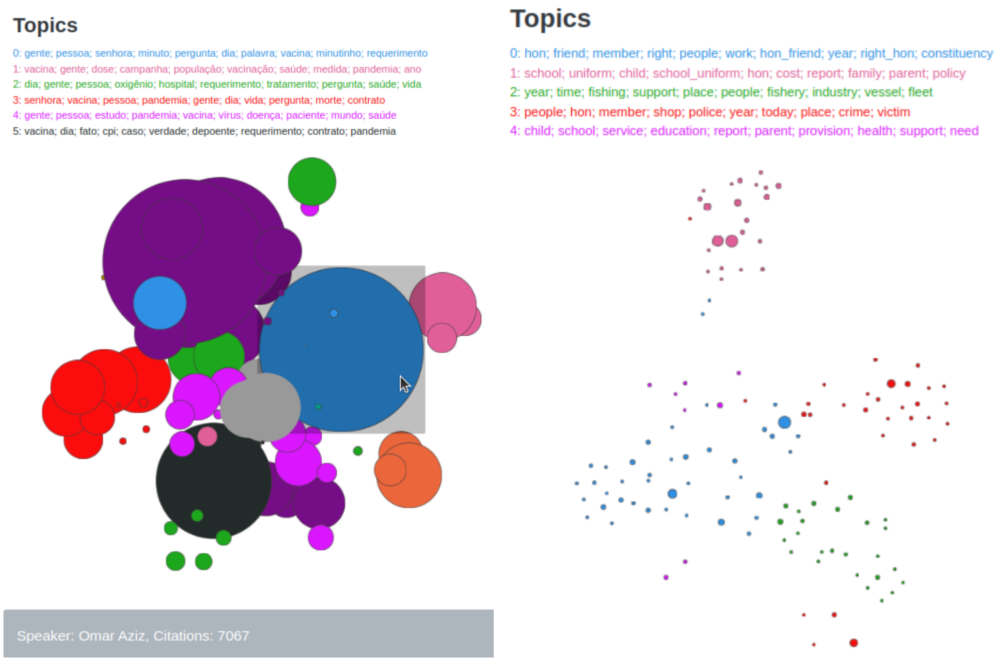
This section presents two case studies using different datasets containing government speeches. We evaluate GoViz on its qualitative aspect, considering two datasets, one in Portuguese and the other in English. The dataset in Portuguese describes the Parliamentary Commission of Inquiry (CPI) dataset about the COVID-19 pandemic. On the other hand, the dataset in English is the ParlSpeech V2 (Rauh & Schwalbach, 2020). Our goal is to show how GoViz can bring the main topics into discussion in both datasets. To validate our goal, we read some of the transcriptions and compared them to what GoViz brings as a result.

4.1. Datasets

The dataset in Portuguese contains the transcriptions from every speech made during the Parliamentary Commission of Inquiry (CPI) of the COVID-19 pandemic in Brazil (see Section 4.2). The dataset was originally made publicly available by the Brazilian federal senate. Later, it was treated and made available by *Base dos Dados*[12], a non-governmental, non-profit, and open-source organization. Besides the speech transcriptions, the entire dataset contains information about the speaker (name, genre, and category; political party acronym; federal unity of Brazil; and the parliamentary bloc) and the speech (besides the text, it has the duration, start, and end time). However, our approach uses only the transcription and the speaker name columns. The transcription column contains a total of 735k words.

The English language dataset we tested is the ParlSpeech V2 (Rauh & Schwalbach, 2020) (see Section 4.3). It contains transcriptions from parliamentary speeches from Austria, the Czech Republic, Germany, Denmark, the Netherlands, New Zealand, Spain, Sweden, and the United Kingdom. For our case, we consider only the House of Commons corpus, which contains data from the United Kingdom. This corpus contains information from 1988 to 2019 about the speech's date, speaker, transcription, political party, and agenda topic. However, for our experiments, we only select data from a single day, the 5^{th} of November 2019, which consists of 114k words.

[12] https://basedosdados.org/dataset/br-senado-cpipandemia



(a) PCI Covid dataset (b) House of Commons corpus

Fig. 6 – t-SNE visualization

4.2. Case Study 1: Parliamentary Commission of Inquiry

A Parliamentary Commission of Inquiry (PCI) often investigates complaints of irregularities carried out by agencies related to the Brazilian government. In 2021, a PCI was instituted to investigate whether there were any crimes of responsibility, possible omissions, and irregularities during the actions promoted in managing the COVID-19 pandemic in Brazil. The discussion on the PCI focused on adopting treatments without scientific evidence, such as using medications like chloroquine and azithromycin. On the other hand, there were campaigns against vaccination and delays in acquiring vaccines. The non-mandatory use of masks and social distancing was also pointed out.

Before the pandemic, Brazil already suffered from intense political polarization (Juvino Santos et al., 2022). Such polarization is reflected directly in the PCI speeches. To show this polarization, we select two groups of senators: the government's base and the opposition. The Government's base comprises: Ciro Nogueira, Jorginho Mello, Marcos Rogério, and Eduardo Girão. The opposition comprises the senators Omar Aziz, Renan Calheiros, Randolfe Rodrigues, Simone Tebet, Humberto Costa, Eduardo Braga, Otto Alencar, and Tasso Jereissati. We selected these names based on their position during PCI.

The treemap present in Figure 3 shows that for the opposition (blue color), the most frequent terms are: *vacina* (vaccine), *contrato* (contract), *requerimento* (application), and *reunião* (meeting). These terms were expected as PCI investigated if there was any irregularity in the vaccine acquisition.

In scatter text visualization (Figure 4), we can see that the opposition uses terms such as "isolation", "immunization", and "accountability" more frequently. From the Government's base, we can see a more financial appeal, with terms such as "funds", "damage", and "public resources". Moreover, the issue of vaccines has been hotly debated on both sides.

In this PCI, there was an effort to define the people responsible for the potential negligence during the pandemic. The network visualization generated by Cytoscape shows the relation between the deponent and the people they mentioned in their speech. In Figure 5, we selected Jorge Kajuro, a Senator who spoke in PCI. His name is in the center of the image. Names cited by Jorge, such as Natalia and Fabio, are from people invited to depose. Fabio, for example, is a former Secretary of Communication of the federal government and has been investigated for possible irregularities. We can also see that names such as Omar Aziz appear multiple times, as he was the president of this PCI.

Figure 6 shows the main terms of the PCI, grouped by topics of interest. For this visualization, we selected 14 topics. In the figure, we limited the number of topics to six to improve visualization. As we can see, the term *vacina* (vaccine) appears as a topic multiple times because they can be interpreted in different scenarios, such as in a more administrative way in topic 2, or focusing on health in topic 3. Considering the PCI scenario and the identified topics, one can understand that the main focus was on investigating contractual points on vaccine acquisitions (topic 2) and conducting treatments and hospital supplies (topics 3 and 4).

4.3. Case Study 2: House of Commons corpus

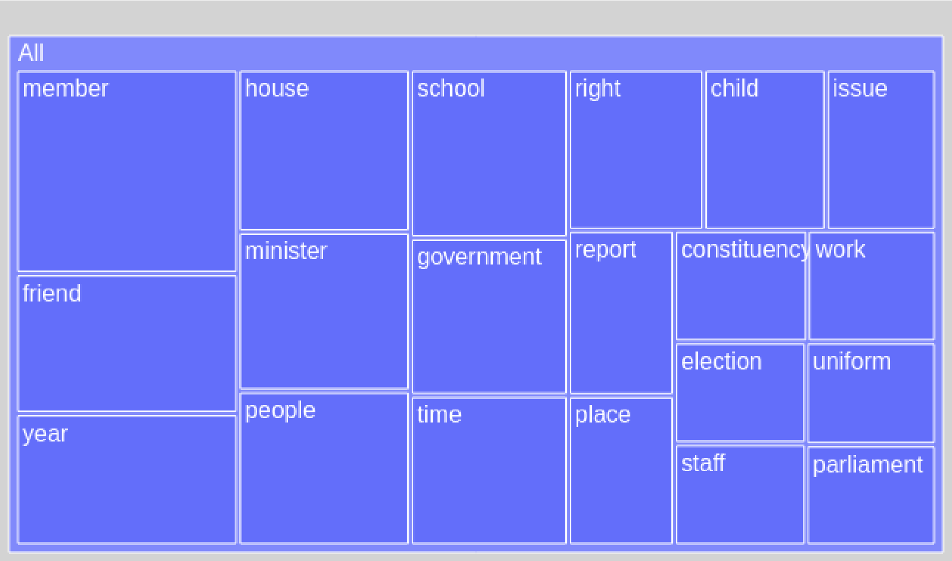


Fig. 7 – Most frequent nouns in House of Commons (ParlSpeech dataset).

Our focus in analyzing the House of Commons corpus was testing whether we could also use GoViz in English documents. The dataset contains information ranging from 1988 to 2019. On this day, the agenda had more than 20 topics involving UK international relations, Human Rights, Climate Change, among others. Thus, considering this scenario without previous knowledge about these topics, we show the most used nouns, without defined groups, in Figure 7. Given the diversified agenda of the day, most terms are more generic. Yet, we can see that some of them indicate the human rights topic as being more frequent, such as 'house', 'right', and 'child'. Figure 6(a) shows the t-SNE visualization for the same day. We can see that the automatically selected topics involve Human Rights and Health (topics 1 and 4), while topic 2 focuses more on the industry. Similar to Figure 6(b), each point in the figure represents a person. The point size indicates the number of times the person mentioned something about the topic (color).

5. Discussion

GoViz is a web-based application that does not require programming skills to use. The main contribution of this work is that it allows extracting information from texts in English or Portuguese without the need to read large amounts of data. With such a feature, GoViz can potentially increase citizens' participation in parliamentary issues. The increase in participation is a consequence of the simplified access to information. In general, our tool brings more interesting results when users provide polarized groups, as shown in Section 4.2. As demonstrated in Section 4, we still obtain an overview of the information presented in the final PCI report without previously knowing the dataset.

Since GoViz is a generic tool capable of handling different datasets, some details were left out. For example, Spacy has excellent accuracy in the named entity recognition task. However, it is not perfect. If we look carefully at Figure 5, we will see terms, such as *respondeu* (replied), *parabéns* (congratulations), and *diretor* (director) that are not person names. Spacy allows customization, but we did not customize it as we expect GoViz to manipulate any dataset, and it should be as generic as possible to deal with a diverse range of topics. We noticed that different names sometimes call the same person throughout the speech. This occurs when the person has a compound name, or the family name is not considered. Jair Messias Bolsonaro (Brazilian president at the time of the PCI) is a good example as he is mentioned throughout the process as *Jair Bolsonaro*, *Jair Messias Bolsonaro*, or just *Bolsonaro*. Although we consider *Jair Bolsonaro* and *Jair Messias Bolsonaro* as the same person when we only have "Bolsonaro", it can also be referring to one of their sons. This makes it almost impossible to guarantee that we always correctly detect all persons' names.

In future work, we aim to incorporate modern models such as Large Language Models and explore sentence representations like Sentence-BERT (Reimers & Gurevych, 2019) to preserve semantic relationships and better capture deep connections within the discourses. Furthermore, to validate GoViz, we aim to conduct user testing to collect feedback and provide design improvements. Other improvements include a new version of GoViz that works entirely online and supports other languages, such as Spanish and French. Finally, we aim to include arc diagrams to improve visualizations and show the correlation among people's dialogues.

6. Conclusion

In this work, we introduce GoViz, a visualization tool that automatically extracts and summarizes speech information. We propose a pipeline that extracts part-of-speech, named-entities, among other information, to provide visualizations that show who the main speakers are, how they connect, and the subject of the speeches. The user can improve GoViz output by adding specific information to the tool, such as person names that belong to two different groups. In this case, GoViz generates visualizations that compare the two groups according to the most spoken terms and connections between them. We evaluate GoViz in both Portuguese and English using open data from the Brazilian government and the United Kingdom. As a result, we show that the generated visualizations can provide insights about the main topics in discussion, as well as the main actors. This is an advance compared to existing tools, which require user input and previous knowledge to work properly.

Acknowledgement

Manssour would like to thank the financial support of the CNPq (N°303208/2023-6) and Griebler thanks the financial support of FAPERGS 09/2023 PqG (N° 24/2551-0001400-4) and CNPq Research Program (N°306511/2021-5)

References

Bernardes, C. B., & Bandeira, C. L. (2016). Information vs Engagement in parliamentary websites – a case study of Brazil and the UK. *Revista de Sociologia e Política*, *24*(59), 91–107. DOI: https://doi.org/10.1590/1678-987316245905.

Bönisch, K., Abrami, G., Wehnert, S., & Mehler, A. (2023). Bundestag-mine: Natural language processing for extracting key information from government documents. *Frontiers in Artificial Intelligence and Applications*, *379*, 391–394. DOI: https://doi.org/10.3233/FAIA230996.

Cantador, I., Viejo-Tardío, J., Cortés-Cediel, M. E., & Rodríguez Bolívar, M. P. (2021). A chatbot for searching and exploring open data: Implementation and evaluation in e-government. *DG.O2021: The 22nd Annual International Conference on Digital Government Research*, 168–179. DOI: https://doi.org/10.1145/3463677.3463681.

Diaz, G. A., Chesñevar, C. I., Estevez, E., & Maguitman, A. (2022). Stance trees: A novel approach for assessing politically polarized issues in twitter. *Proceedings of the 15th International Conference on Theory and Practice of Electronic Governance*, 19–24. DOI: https://doi.org/10.1145/3560107.3560296.

Harrison, T. M., Guerrero, S., Burke, G. B., Cook, M., Cresswell, A., Helbig, N., Hrdinová, J., & Pardo, T. (2011). Open government and e-government: Democratic challenges from a public value perspective. *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times*, 245–253. DOI: https://doi.org/10.1145/2037556.2037597.

Heer, J., Bostock, M., & Ogievetsky, V. (2010). A tour through the visualization zoo: A survey of powerful visualization techniques, from the obvious to the obscure. *Queue*, *8*(5), 20–30. DOI: https://doi.org/10.1145/1794514.1805128.

Hinton, G. E., & Roweis, S. (2002). Stochastic neighbor embedding. *Advances in neural information processing systems*, *15*, 857–864.

Inter-Parliamentary Union. (2022). *Global Parliamentary Report 2022 - Public engagement in the work of parliament* (2022nd ed.) [publisher: Inter-Parliamentary Union].

Juvino Santos, L. R., Balby Marinho, L., & Calazans Campelo, C. E. (2022). Uniting politics and pandemic: A social network analysis on the covid parliamentary commission of inquiry in brazil. *Proceedings of the Brazilian Symposium on Multimedia and the Web*, 99–107. DOI: https://doi.org/10.1145/3539637.3556992.

Keim, D. A., & Oelke, D. (2007). Literature fingerprinting: A new method for visual literary analysis. *2007 IEEE Symposium on Visual Analytics Science and Technology*, 115–122. DOI: https://doi.org/10.1109/VAST.2007.4389004.

Kessler, J. (2017). Scattertext: A browser-based tool for visualizing how corpora differ. *Proceedings of ACL 2017, System Demonstrations*, 85–90.

Khartabil, D., Collins, C., Wells, S., Bach, B., & Kennedy, J. (2021). Design and evaluation of visualization techniques to facilitate argument exploration. *Computer Graphics Forum*, *40*(6), 447–465. DOI: https://doi.org/https://doi.org/10.1111/cgf.14389.

Montani, I., Honnibal, M., Honnibal, M., Landeghem, S. V., Boyd, A., Peters, H., McCann, P. O., Samsonov, M., Geovedi, J., O'Regan, J., Altinok, D., Orosz, G., Kristiansen, S. L., de Kok, D., Miranda, L., Roman, Bot, E., Fiedler, L., Howard, G., ... Böing, B. (2020, June). spaCy: Industrial-strength Natural Language Processing in Python. DOI: https://doi.org/10.5281/zenodo.6621076.

Palma, I., Ladeira, M., & Reis, A. C. B. (2021). Machine learning predictive model for the passive transparency at the brazilian ministry of mines and energy. *DG.O2021: The 22nd Annual International Conference on Digital Government Research*, 76–81. DOI: https://doi.org/10.1145/3463677.3463715.

Quintero-Angulo, R. A. D., Sánchez-Torres, J. M., & Cardona-Román, D. M. (2020). Problem areas in e-participation: A systematic review. *Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance*, 544–550. DOI: https://doi.org/10.1145/3428502.3428584.

Rajaraman, A., & Ullman, J. D. (2011). Data mining. In *Mining of massive datasets* (pp. 1–17). Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139058452.002.

Rauh, C., & Schwalbach, J. (2020). The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies. DOI: https://doi.org/10.7910/DVN/L4OAKN.

Rehurek, R., & Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, *3*(2), 45–50.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 3980–3990. https://arxiv.org/abs/1908.10084

Steinbauer, M., Hiesmair, M., & Anderst-Kotsis, G. (2016). Making computers understand coalition and opposition in parliamentary democracy. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9820 LNCS*, 265–276. DOI: https://doi.org/10.1007/978-3-319-44421-5_21.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(11), 2579–2605.

Volkovskii, D., & Filatova, O. (2022). Agreement and disagreement in american social media discussions (evidence from facebook discussions on the second impeachment of d. trump). *Proceedings of the 15th International Conference on Theory and Practice of Electronic Governance*, 221–228. DOI: https://doi.org/10.1145/3560107.3560144.

Ward, M., Grinstein, G., & Keim, D. (2015). *Interactive data visualization: Foundations, techniques, and applications, second edition*. CRC Press.

Wattenberg, M., & Viégas, F. B. (2008). The word tree, an interactive visual concordance. *IEEE Transactions on Visualization and Computer Graphics*, *14*(6), 1221–1228. DOI: https://doi.org/10.1109/TVCG.2008.172.