Mining United Nations General Assembly Debates
Final report for NLP Course, Winter 2023

Mateusz Grzyb and **Mateusz Krzyziński** and **Bartłomiej Sobieski** and **Mikołaj Spytek**
Warsaw University of Technology
{mateusz.grzyb3, mateusz.krzyzinski,
bartlomiej.sobieski, mikolaj.spytek}.stud@pw.edu.pl

supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Abstract

The United Nations General Assembly (UNGA), a vital hub for international diplomacy, convenes annually to address pressing global issues, generating a vast repository of transcripts dating back to 1946. This project aimed to construct a comprehensive dataset, enriching these transcripts with metadata, enabling a deeper understanding of the shifts in international diplomacy and global priorities over time. This vast corpus presents tremendous challenges in terms of volume and complexity. To overcome these challenges, we leveraged Natural Language Processing (NLP) techniques to extract meaningful insights from these transcripts, complemented by the collected metadata and statistical text features. This endeavor is underpinned by the growing importance of NLP and text mining in social and political sciences, emphasizing the relevance of the UNGA corpus in these fields. It also explores the practical application of topic modeling, particularly the state-of-the-art transformer-based BERTopic model. Our comprehensive approach encompasses tracking evolving topics, examining speech characteristics, and convincingly visualizing the results, all with the goal of decidedly answering prominent research questions about UNGA statements.

1 Introduction

The United Nations[1] (UN) is an intergovernmental organization established in 1945 after World War II in an effort to prevent any future global conflicts. At its formation, it consisted of 51 member states; as of 2023, it has 193. Its primary goals are maintaining international peace and security, protecting widespread respect for human rights and promoting friendly cooperation among all nations.

The General Assembly[2] (GA) is the central policy-making and representative organ of the UN. It takes place in regular yearly sessions and gathers all UN members. During the first week of each new session, a so-called General Debate is held. It is a high-level event which gives the appointed delegates an opportunity to bring to attention issues most important for the member states they represent.

Transcripts from all such debates, beginning from 1946 onward, are publicly available. They constitute a valuable source of information regarding the changing dynamic of contemporary international relations. Therefore, their analysis is relevant both from the point of view of ordinary citizens, whose interests ought to be well represented, and political scientists, whose research should be supported by a close observation of reality.

However, the large number and volume of the source material described render its meticulous manual analysis almost impossible. Fortunately, methods of computer-based natural language processing (NLP), which are currently developing at a rapid peace, enable the automatic extraction of de-

[1] https://www.un.org/en/about-us
[2] https://www.un.org/en/ga/about/background.shtml

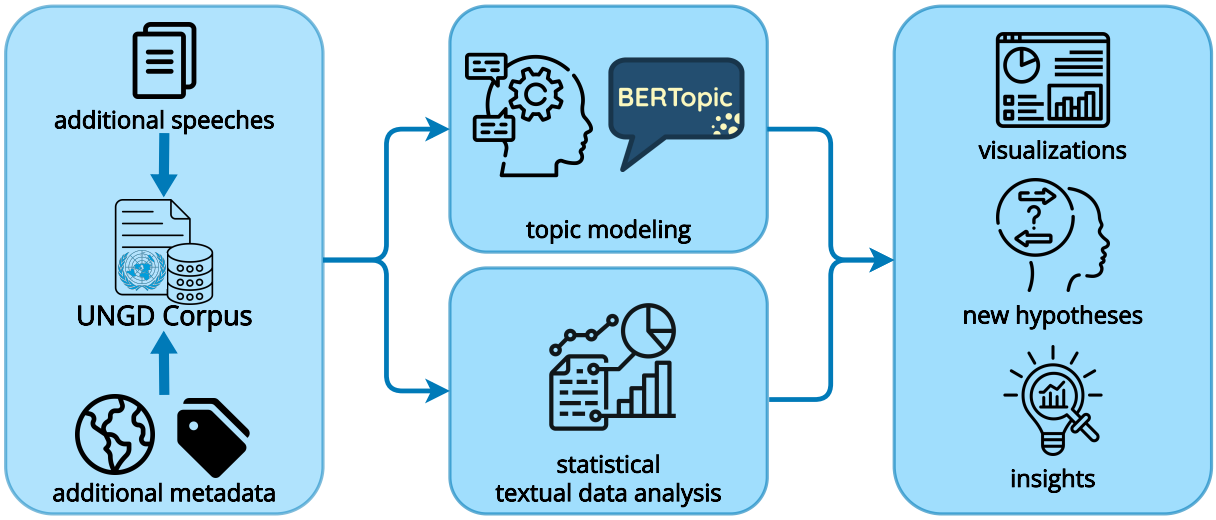

Figure 1: Diagram showcasing main steps of the project.

tailed information and complex relationships from massive amounts of text data. These methods are also gaining popularity in application to political science, which highlights an excellent opportunity regarding the aforementioned data.

In this work, we use the NLP methods to explore and analyze the UNGA data after supplementing it with transcripts from the newest debates and enriching it with new metadata. In this way, we want to provide way to answer meaningful questions, such as: What lexical and statistical features characterize the statements under consideration? What topics and themes are being addressed during the debates? Do these factors depend on the time of the debate and the state represented by the speaker?

The overarching goals of this project can be summarized as follows:

- Preparing a complete dataset of statements presented at the UNGA in the years 1946-2023, complete with metadata concerning the country, date, name, role of the speaker, and enriching this metadata by additional features from external sources.

- Exploring the gathered corpus, along with collected metadata and newly calculated speech statistics and preparing visualizations of the results.

- Applying a state-of-the-art topic modeling techniques based on transformers to extract evolution of themes present at the general debates and appropriate aggregation of the results.

Topic modeling and statistical analysis results can be visualized in the web application to generate answers to research questions, insights and new hypotheses. The overview of the project is presented in Figure 1.

2 Related work

2.1 Text mining and analysis of UN General Debates

The application of text mining methods in the fields of social science and political science has been gaining significant popularity due to their ability to efficiently process vast amounts of textual data (Hollibaugh, 2018). This growing interest can be attributed to the utilization of modern Natural Language Processing (NLP) tools for addressing various challenges in these domains (Nay, 2018; Glavaš et al., 2019).

This interest extends to the analysis of United Nations General Debates, primarily due to the substantial coverage of major global issues within these deliberations. (Baturo et al., 2017) emphasized that these valuable resources had been overlooked for many years and took the initiative to create the initial version of the corpora, named `UNGDC`. It comprised over 7,300 country statements from 1970 to 2014. Their pioneering work involved the application of statistical linguistic methods, such as wordscores (Laver et al., 2003) and correspondence analysis (Benzecri,

1992), showcasing how the UNGDC could be employed to reveal single and multiple dimensions of government preferences.

Furthermore, in more recent developments, the UNGDC has been updated and made publicly available (Jankin et al., 2017). The extended version of the corpus now encompasses data from 1946 to 2022, featuring over 10,000 speeches from representatives of more than 193 countries with additional metadata as shown in Table 1. This comprehensive collection of global political discourse stands as one of the most extensive resources of its kind. In their recent study (Dasandi et al., 2023), the authors provided additional examples of how such corpus can be leveraged, including the utilization of topic modeling techniques to explore countries' engagement with sustainable development goals. Building upon their research, we aim to further enrich the created corpora and expand the application of topic modeling techniques, extending the scope of analysis beyond their study.

2.2 Topic modeling in social science

One of the most commonly employed NLP methods in social science is topic modeling (Vayansky and Kumar, 2020). It facilitates the automatic identification of latent topics within extensive text collections. In the realm of social sciences, topic modeling serves as a valuable tool for tasks like data exploration and the quantitative analysis of text data that may be challenging to objectively measure otherwise (Valdez et al., 2018). Recently, researchers have started exploring its applications beyond explanatory purposes. As suggested by Valdez et al. (2018), topic modeling can also be harnessed to compare structured corpora, enabling the investigation of semantic similarities and differences, which aligns with the focus of our study.

Various variations of topic modeling have been employed to social science analyses. For instance, Grimmer (2010) introduced the Bayesian Hierarchical Topic Model, a probabilistic model designed to capture hierarchical relationships among topics, which he used to analyze press releases from American senators.

However, it is more common to use already existing models by applying them to new datasets. For example, Greene and Cross (2017) employed a method based on two layers of Non-negative Matrix Factorization (NMF) (Lee and Seung, 1999) to dynamically explore the content of speeches delivered by members of the European Parliament. NMF, typically used in dimensionality reduction, is adapted in this context to uncover underlying topics based on factorization of a term-document matrix into two non-negative matrices, where one matrix represents topics and the other matrix represents the weights of topics in documents.

Finally, Żółkowski et al. (2022) used a widely used topic modeling technique Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to compare national energy and climate plans established by 27 Member States of the European Union. They used LDA due to its great simplicity and interpretability, resulting from the presentation of the documents as a mixture of melt, which was used in their analyses.

More recently, state-of-the-art topic modeling techniques based on transformers have gained increasing traction and have demonstrated their effectiveness in various applications. The most prominent example is BERTopic (Grootendorst, 2022) that has already been used in many fields and is currently considered state-of-the-art . Notable examples include the analysis of public sentiment on the Internet during the monkeypox outbreak (Ng et al., 2022), topic extraction from financial policies (Clapham et al., 2022), or analysis of news impact on financial markets (Chen et al., 2023). Our work will also make use of the capabilities offered by this approach due to its modular nature that allows for substituting different internal components and state-of-the-art capabilities. As of today, there are no other frameworks based on deep learning methods that would allow for a comparative analysis. Considering that and the nature of our task, we will focus on applying the existing methods rather than performing a benchmark study.

Specialized tools simplify similar analyses, with several comprehensive frameworks streamlining the entire modeling process, from training to result evaluation. Notable examples include the Gensim framework (Rehurek and Sojka, 2011), Topic Modeling API (ToModAPI) (Lisena et al., 2020), Optimizing and Comparing Topic Models is Simple (OCTIS) (Terragni et al., 2021), or Interactive Topic Model Trainer (ITMT) (Calvo Bartolomé et al., 2023). Additionally, tools like Homologous Automated Document Exploration and Summarization (HADES) (Wilczyński et al., 2023) facilitate the comparative analysis and com-

Table 1: Excerpt of the available metadata from the original UNGDC (foundation corpus).

Year	Session	ISO Code	Country	Name of Person Speaking	Post
2004	59	ZWE	Zimbabwe	Mr. Robert Mugabe	President
2003	58	AFG	Afghanistan	Hâmid Karzai	President

parison of structured corpora. However, in our work, we will make use of the capabilities offered by the BERTopic package, leveraging its specific functionalities and extensions to maximize its potential.

3 Approach & Methodology

3.1 Dataset preperation

Our dataset is built upon the foundation of the existing UNGDC dataset, a publicly available resource (Jankin et al., 2017; Dasandi et al., 2023). Encompassing the period from 1946 to 2022, this corpus comprises over 10,000 speeches from representatives of more than 193 countries, each speech stored in individual `.txt` files. Accompanying this wealth of textual data are basic metadata, structured as outlined in Table 1.

The creation of this corpus involved overcoming substantial challenges, particularly in processing historical data. The data preceding 1992, sourced from the United Nations Digital Library[3], existed as image copies with varying structures across years. To address this, the corpus's authors manually extracted text, supported by the use of optical character recognition software.

Initially, we updated it by integrating newly delivered speeches from 2023, drawn from country-specific templates on the UN General Assembly website, as these speeches are not consolidated on the United Nations Digital Library.

Addressing data quality concerns, we correct found errors and inconsistencies, particularly in the metadata. Moreover, our endeavor involves enriching the dataset's metadata by introducing additional variables and indices that provide a more comprehensive understanding of characteristics of countries and their situation across different years. Furthermore, we integrate information from the United Nations geoscheme, providing division into regions and sub-regions. This comprehensive approach ensures a robust and multi-dimensional characterization of countries, enriching the dataset for a more in-depth analysis of UNGA debates.

[3] https://digitallibrary.un.org

3.2 Data preprocessing

The first stage of data preparation for analysis is applying a set of preprocessing steps to the prepared corpus, which is the basic procedure in the statistical analysis of data of text modality (Manning et al., 2010). For this purpose, we use the tools and models available in the Python spaCy package (Honnibal et al., 2020).

Following the tokenization and lemmatization procedures, we further refine the dataset by excluding stopwords. To achieve this, we utilize an existing, ready-made list for the English language. These preprocessing steps are essential in streamlining the subsequent analyses, ensuring the focus remains on meaningful content while eliminating unnecessary elements.

3.3 Statistical textual data analysis

Simple statistical analysis of the text corpora can be used to extract insightful information without sophisticated modeling methods. Our approach involves employing intuitive and well-established methods to extract meaningful insights. For each speech document, we generate a list of the most frequent words along with their counts, alongside a diverse array of descriptive statistics. These include metrics such as the number of tokens, unique tokens, sentences, and characters. Additionally, we compute statistical measures such as the mean, median, and standard deviation of token length and sentence length.

To delve deeper into the analysis, we incorporate more sophisticated approaches and text statistics. We adapt the lexical dispersion plot proposed in the quanteda framework (Benoit et al., 2018), providing a graphical representation of word distribution within the speeches. Furthermore, we leverage measures from the TextDescriptives library (Hansen et al., 2023), encompassing various readability metrics, part-of-speech proportions, and document coherence values. In total, this comprehensive approach yields over 60 distinct statistics for each speech text.

We leverage the calculated values by creating diverse visualizations that depict trends over time and showcase relationships between different statistical features. In the context of mining General Assembly debates, such analyses prove invaluable for examining the evolution of changes in speeches from specific countries over time and facilitating comparisons between documents from different regions or countries.

3.4 BERTopic

Theoretical setup. BERTopic is a modular framework designed for topic modeling implemented in Python. It was built around interchangeable components to allow for performing versatile experiments. By default, BERTopic uses a pretrained Large Language Model, which inspects semantic similarity between the words contained in the document. It was first proposed in (Grootendorst, 2022) and due to the well-documented open-source Python implementation, it has gained popularity in the NLP community. It is important to stress that BERTopic uses word embeddings to generate topics in the article so semantic similarity of documents plays a crucial role in discovering meaningful topics.

The method of extracting topics from a corpus of documents can be summarized with four main steps.

1. The documents are converted to their embedding representations using the BERT (Devlin et al., 2019) pre-trained Large Language Model (by default). Using other Language models is possible, and can substantially change the quality of the selected topics.

2. The embeddings are processed with the UMAP (McInnes et al., 2018) dimensionality reduction model to improve clustering results.

3. Clustering analysis is performed on the reduced embeddings using the HDBSCAN (McInnes et al., 2017) algorithm.

4. Human-readable descriptions of the topics are generated by using the class-based TF-IDF (or *c-TF-IDF* for short) technique (Salton and Buckley, 1988) on each cluster separately to extract the most meaningful words from all topics.

BERTopic assigns only one topic per document as the underlying assignment is done via HDBSCAN, which assigns one document to just one cluster.

In the main process of clustering, BERTopic operates on uninterpretable word embeddings and the human-understandable descriptions are extracted at the last step using the TF-IDF technique for each cluster. For each word in each document, a metric is calculated, and the words with the highest scores are chosen as topic descriptions.

Drawbacks and limitations. The BERTopic model does not allow for a manual selection of the desired number of topics in the corpus. It generates as many topics, as there were clusters selected by the HDBSCAN method which can be indirectly tuned with some hyperparameters. Additionally, each document is only assigned one topic, instead of a mixture of topics as in popular older approaches e.g., (Lee and Seung, 1999; Blei et al., 2003). A relatively high computational cost is another disadvantage of this method, as the usage of a pretrained Large Language Model for acquiring embeddings for each document is time-consuming.

BERTopic extensions. The described basic BERTopic framework lends itself to many extensions making it more suitable for answering the research questions stated for this project. One such extension is its adaptation to the dynamic topic modeling framework first proposed by Blei and Lafferty (2006), which allows for analysing the evolution of topics over time.

To achieve this, first, BERTopic is fitted to the entire corpus as if there were no temporal aspect in the documents. This produces a general topic model – a global representation of general topics spanning the documents. Next, for each selected time point (UNGA Session) and each general topic, a separate TF-IDF representation is created. This allows us to follow particular topics through the years and examine how they evolve, and how the way they are spoken about changes. Lastly, these specific model representations can be further fine-tuned to either emphasise the global nature of these topics or to focus on their evolution over time.

Another such extension is semi-supervised topic modeling. This approach allows for an efficient usage of the additional metadata coming

from the highly structured dataset such as the country of the speaker, the general region of their country or even the year of the speech. By providing the metadata along with the text of the documents, we steer the dimensionality reduction (using supervised or semi-supervised UMAP) of the embeddings into a space that better follows the relationships between documents.

BERTopic visualizations. The used Python implementation comes with a wide array of tools for topic visualisation. These allow for depicting relative similarity between topics, the relationship between documents and topics, the hierarchical structure of the generated topics as well as visualizations which take the temporal aspect of the analysis into account. Additionally, the topic data is stored in an easily accessible format, which facilitates the creation of other kinds of visualizations.

3.5 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a basic topic model, often used as a baseline for other advanced methods. This technique assumes that each document is a probability mixture of many *latent* (hidden, unseen) topics and in turn, each topic is a probability mixture of the words from the vocabulary.

The procedure for obtaining topics involves iterative steps of estimating the distribution of topics in each document and the distribution of words in each topic. The algorithm begins by assigning random topics to words in the documents. Subsequently, a process of iteratively updating the topic assignments based on the observed words and refining the topic distributions occurs, with parameters estimated using Bayesian inference techniques. This iterative process converges towards a stable representation of topics and their associated word distributions across the entire document corpus.

Despite its popularity, the LDA approach has many limitations. Due to its simplicity, it can only find a fixed number of topics. Another drawback is the lack of support for correlated topics. Similar topics with small differences are most often clumped into bigger topics which leads to a loss of detail. Furthermore, the generated topics are static – there is no possibility of adding a new topic to a fitted model by supplying more documents. All of these limitations are addressed in the state-of-the-art BERTopic method.

3.6 Technical considerations

The majority of the project uses the Python programming language. The required versions of libraries needed to reproduce the results of this project are listed in the `requirements.txt` file attached to the source code. All computations have been conducted locally, without using any HPC clusters or cloud solutions.

The whole framework produced as a result of this project is available as a Docker container image at `https://hub.docker.com/r/krzyzinskim/nlp-unga-debates`. This allows for reproducibility and easily running the application on any hardware. Additionally, the application is available online at `https://nlp-unga-debates-g2o6gnzttq-lm.a.run.app/`.

4 Experiments & Results

4.1 Enhanced UNGD corpus

The result of our work is the enhanced version of the United Nations General Debate Corpus, now inclusive of speeches delivered in 2023. The incorporation of these speeches posed challenges due to their unavailability in the United Nations Digital Library, necessitating extraction from national templates, each unique to its respective country or, in some instances, absent. Consequently, the corpus comprises 120 speeches, constituting a subset of the 195 originally delivered. In total, the enhanced UNGDC consists of 10,679 speeches.

The enhancement extends to improved metadata integrity, achieved through an iterative feature engineering process. This process involved various refinements, such as rectifying typographical errors (e.g., transforming *Trinidad and Tobado* to *Trinidad and Tobago* and *Swistzerland* to *Switzerland*), harmonizing nomenclature to standardize varied representations (e.g., *Holy SEE*, *Vatican*, *Holy See*, and *Vatican City State* into a singular *Vatican*), and addressing inaccuracies in ISO codes (e.g., transforming *POR* to *PRT* for Portugal). Further data engineering and analyses were facilitated by resolving complications related to countries that no longer exist. For instance, for countries with current equivalents, the current code was employed (e.g., Ukrainian SSR and Ukraine have *UKR* code). In cases of coun-

Table 2: Newly added variables to the metadata.

Column name	Description	Source	Coverage (%)
Population	the number of people	Gapminder	99.88
TFR	Total Fertility Rate, average number of children per woman	Gapminder	96.87
HDI	Human Development Index, composite index of life expectancy, education and per capita income indicators	United Nations	50.92
GDP	Gross Domestic Product, at constant 2015 prices in US dollars	World Bank	80.61
Unemployment Rate	percentage of the labor force (15 years+) without jobs	United Nations	52.17
Gini	Gini coefficient, measure of income inequality	Gapminder	97.67
CO2	carbon dioxide emissions from fossil fuels, in tonnes per capita	Global Carbon Project	96.34
Democracy Index	index measuring the quality of democracy	Our World in Data	23.23
Region Name	6 continental regions derived by United Nations Statistics Division (UNSD)	United Nations	100.00
Sub-region Name	22 sub-regions derived by United Nations Statistics Division (UNSD)	United Nations	99.88

tries that have split, their former codes were used (e.g., Yugoslavia – YUG), with metadata alignment achieved through pertinent operations on the data pertaining to individual components of the erstwhile state, where applicable. Finally, the letter casing is fixed and consistent for the names of speakers and their positions. In total, 266 ISO codes, 1,862 country names, 3,877 names of speakers, and 3,199 of their positions were improved for enhanced accuracy.

A significant augmentation in the new version of the UNGD corpus is the inclusion of additional metadata, comprising 10 new covariates matched to all speeches whenever applicable. The detailed description of this data, along with its sources and coverage (percentage of documents for which they were matched), is delineated in Table 2. The data were drawn from diverse sources, including Gapminder[4], Our World in Data[5], and the United Nations[6] alongside its agencies and programs. It is essential to acknowledge that certain covariates, such as HDI and Democracy Index, represent metrics that had not yet been introduced in 1946, contributing to inherent missing data.

4.2 Comparative evaluation of topic models

To justify the choice of BERTopic as the main focus for this project, we compare it using different metrics with the LDA method. LDA is considered to be a simple baseline for all topic modeling tasks. Additionally, in this step, we evaluate different versions of the BERTopic, with different methods of embedding the documents. In addition to proving the superiority of BERTopic over the baseline, it also allows us to select the best document embedding method for further study.

Document embeddings. Creating a numerical representation of the documents is the first key step in the BERTopic framework. The clustering step, which actually extracts topics works in this embedding space, so quality embeddings have a great influence on the overall quality of the topics. However, there is not one perfect embedding model and different methods suit different use cases. We evaluate BERTopic with different embedding methods to quantitatively select the best one for creating topic analyses.

First of all, we use three sentence transformer models (Reimers and Gurevych, 2019) for obtaining embeddings: the smallest and fastest `all-MiniLM-L6-v2` with 6 layers, the larger `all-MiniLM-L12-v2` with 12 layers, and the largest `all-mpnet-base-v2` with an enhanced corpus used for training. Additionally, we consider embeddings extracted from the RoBERTa (`roberta`) (Liu et al., 2019) and DistilBERT (`distilbert`) (Sanh et al., 2019) models. These have been chosen as they are recommended by the authors of the BERTopic Python library for general topic modeling.

Using these embeddings, we perform the rest of the topic modeling steps in the BERTopic framework and for each of them, we obtain a separate set of topics. Additionally, we consider topics ob-

[4] https://www.gapminder.org/data/
[5] https://ourworldindata.org
[6] http://data.un.org/datamartinfo.aspx

Table 3: Topic coherence metric values.

Number of topics	Topic modeling method					
	LDA	BERTopic				
		all-MiniLM-L6-v2	all-MiniLM-L12-v2	all-mpnet-base-v2	distilbert	roberta
10	0.387	0.458	0.432	0.433	**0.463**	0.324
20	0.397	0.450	0.422	0.427	**0.499**	0.324
50	0.401	0.442	0.428	0.456	**0.493**	0.324

Table 4: Topic diversity metric values.

Number of topics	Topic modeling method					
	LDA	BERTopic				
		all-MiniLM-L6-v2	all-MiniLM-L12-v2	all-mpnet-base-v2	distilbert	roberta
10	0.250	**0.630**	0.500	0.560	0.570	0.567
20	0.250	0.445	0.430	0.450	**0.470**	0.567
50	0.258	0.358	0.342	**0.390**	0.372	0.567

tained by the LDA method (with $n = 10$, $n = 20$ and $n = 50$ topics) and evaluate them using established metrics.

Metrics. The following metrics have been used for evaluating the topic models:

1. **Topic coherence** (Lau et al., 2014) metric utilizes various statistics drawn from the reference corpus to evaluate how well the extracted topics are 'supported' by it. In other words, this measure indicates the degree of 'interpretability' of the obtained topics in context of their source. The calculation of this metric is a complex, multi-step process and was described in detail in Röder et al. (2015). Specifically, we use the C_v version of this metric, which achieved the highest correlation with human ratings.

2. **Topic diversity** (Dieng et al., 2020) In contrast to topic coherence, the topic diversity is a much simpler metric calculated based solely on the extracted topics and evaluating how much variability there is among them. It is defined as the ratio of the number of unique words among top k words of each topic and the upper bound of this value, which equals the product of the number of words considered k and the total number of topics.

Results. Table 3 shows values of the coherence metric and Table 4 shows values of the topic diversity metric, all corresponding to a given topic modeling method and number of topics combination. It should be noted that values for the `roberta` method should be ignored due to it always extracting only 6 distinct topics.

By both metrics, the LDA method performs significantly worse than methods from the BERTopic family. Regarding the topic coherence metric, the `distilbert` method is by far the best for any number of topics. In terms of topic diversity metrics, the `distilbert` method is the first or second best depending on the number of topics. For this reason, we decide to utilize the `distilbert` method for our further analyses.

4.3 Interactive application

To address the challenges posed by the substantial size and enriched nature of the created corpus, an interactive application has been developed using Streamlit data app framework. Accessible via the web at `https://nlp-unga-debates-g2o6gnzttq-lm.a.run.app/`, this application serves as a user-friendly tool, facilitating dynamic exploration and analysis of the new UNGD corpus with collected metadata and statistics.

The application is designed to allow users, including political science researchers, to navigate through the corpus more efficiently, and conduct analyses tailored to specific dimensions of interest. The app is structured to allow users to customize their analyses, selecting specific timeframes, countries, or variables. Results are predominantly presented through visualizations, enhancing the interpretability of results and facilitating the communication of findings.

The application is structured into four distinct sub-pages, each presenting different facets of the

corpora. Results visible in different sub-pages are illustrated in the Figures in Appendix B. The first sub-page, *Speech Viewer*, enables users to explore individual speeches selected by year and country. It displays the speech text alongside relevant metadata and basic analyses, including the most common words and a lexical dispersion plot (see Figure 3). Another tab, *Speech Comparer*, allows for a side-by-side comparison of two speeches.

The *Analysis Over Years* tab features visualizations that illustrate analyses over the selected time period for chosen groups of countries. Users can explore trends in the most common words and observe how measures change over time. For instance, the period following a war, 1946-1955, features common words like *peace*, *war*, *Charter*, *Soviet*, and *USA*, reflecting the aftermath and geopolitical landscape of that era. Conversely, in the aftermath of the 9/11 World Trade Center attacks (2001-2010), the discourse evolves to include words such as *development*, *security*, *peace*, *community*, and *terrorism*. Such analyses reflect the evolving themes and priorities in the UNGA debates over different epochs. Additionally, Figure 4 depicts a noteworthy increase in the fraction of unique tokens over time, suggesting a broader spectrum of diverse issues discussed in these debates.

Finally, the *Speech Attributes* offers a platform for in-depth exploration of the collected features. Users can compare values of various selected variables across different years and countries, thereby gaining a multifaceted perspective on the interplay between linguistic elements and contextual attributes. Moreover, this tab facilitates an examination of how the frequency of a chosen word correlates with metadata variables describing the countries. For example, one can observe that countries with higher democracy index seem to have more readable texts of speeches (see Figure 5) or the countries with high CO2 emissions per capita tend to not speak a lot about climate issues (see Figure 6).

Supplementing these four sub-pages, the application incorporates an additional section specifically focused on analysis using topic modeling, which is elaborated upon in more detail in the subsequent section.

4.4 BERTopic analyses

We perform advanced analysis using the BERTopic package to highlight different facets of the obtained topic selection. In the next sections, the theoretical aspects of the analysis method are described, along with examples on the UNGD corpus.

Intertopic distance map. The simplest visualization of the obtained topic selection is the Intertopic distance map. This is a two-dimensional projection, which allows to compare the relative sizes and distances between the topics. The technique is inspired by LDAvis (Sievert and Shirley, 2014). It works by embedding the c-TF-IDF representations in 2 dimensions using UMAP (McInnes et al., 2018).

In this visualization each topic is represented by a circle, whose size is determined by the number of documents assigned to the topic and the distance between circles corresponds to the distance between topic embeddings. By mousing over a topic users can find out which words describe a given topic, and how many documents belong to it.

Similarity matrix. The c-TF-IDF embeddings of topics can be also compared quantitatively, not only visually. That is done by measuring the cosine similarity between pairs of topics. We then obtain a numerical similarity score between each pair of topics. These scores can be graphically represented in a similarity matrix as a complement of the Intertopic distance map.

Hierarchical topic structure. The topics that created using BERTopic can be hierarchically reduced. In order to understand the potential hierarchical structure of the topics a dendrogram visualization can be used to show clusters and how they relate to one another. This helps select the appropriate number of topics in the corpus, as well as when reducing the number of automatically obtained topics.

Dynamic topic modeling. Dynamic topic modeling (DTM) is a collection of techniques to analyse topics' evolution over time. These methods allow for understanding how a topic is represented across different times. For example, in 1995 people may talk differently about some political aspect than those in 2015. Although the topic itself

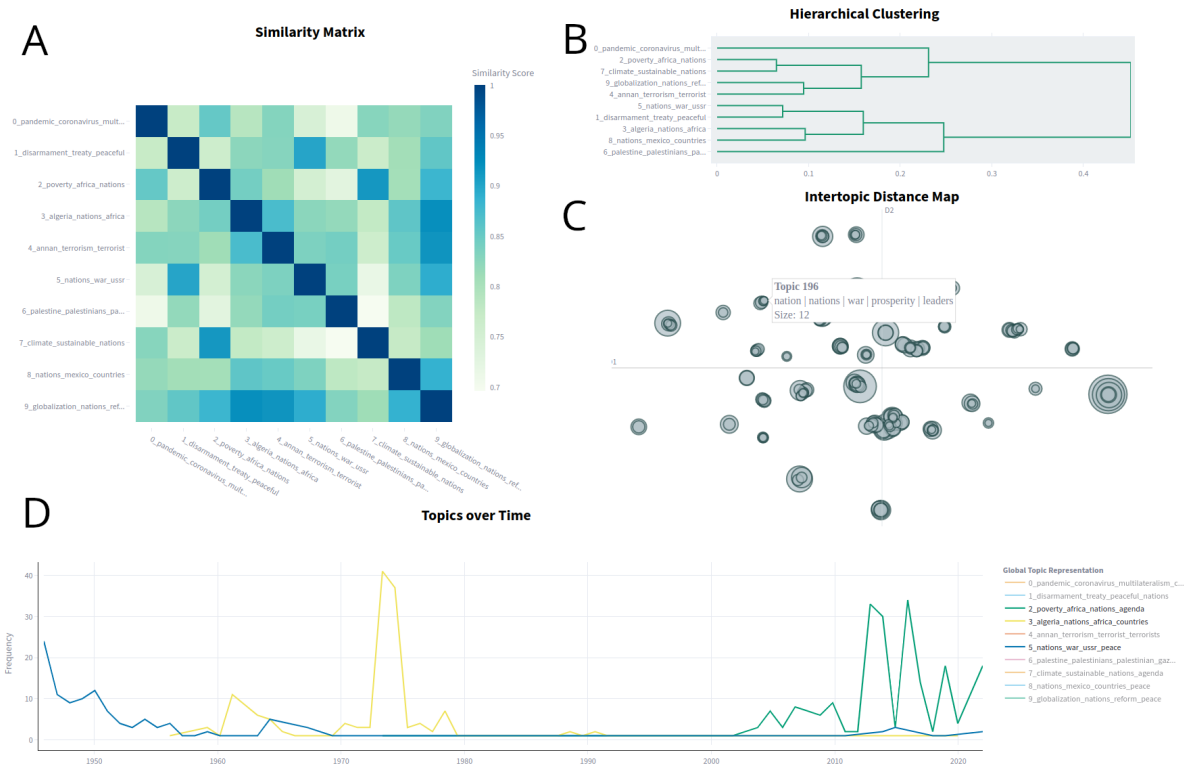

Figure 2: Visualizations obtained using BERTopic on the UNGD corpus. A – similarity matrix; B – Hierarchical clustering; C – Intertopic distance map; D – Topics over time.

remains the same, the exact representation of that topic might differ.

BERTopic allows for DTM by calculating the topic representation at each timestep without the need to run the entire model several times. To do this, we first fit BERTopic as if there were no temporal aspect in the data. Thus, a general topic model will be created. We use the global representation to find the main topics that can be discussed differently at different timesteps. For each topic and timestep, we calculate the c-TF-IDF representation. This will result in a specific topic representation at each timestep without the need to create clusters from embeddings as they were already created.

Next, there are two main ways to fine-tune further these specific topic representations, namely globally and evolutionary. A topic representation at timestep t can be fine-tuned globally by averaging its c-TF-IDF representation with that of the global representation. This allows each topic representation to move slightly towards the global representation while keeping some of its specific words.

A topic representation at timestep t can also be fine-tuned evolutionary by averaging its c-TF-IDF representation with that of the c-TF-IDF representation at timestep $t-1$. This is done for each topic representation allowing for the representations to evolve over time. We use both fine-tuning methods, as this is the default choice in the BERTopic package.

For illustrative purposes, we include an example of applying dynamic topic modeling to the UNGD corpus in Figure 2 (D). Importantly, we observe distinct periods of popularity of well-defined topics that agree with the historical timeline of events.

Topics per class. We might also be interested in how certain topics are represented over certain categories. For example, countries from specific regions might present some specific views on the given topic. Instead of running the topic model per class, we can simply create a topic model and then extract, for each topic, its representation per class. This allows us to see how certain topics, calculated over all documents, are represented for certain subgroups.

5 Discussion & Conclusions

Our contributions can be summarized with a set of complementary approaches to

1. **dataset creation.** The UN GA general debates corpus has been carefully sanitized, supplemented with the latest speeches and enriched with over 10 additional geopolitical features.

2. **development of interactive application.** An interactive and containerized application has been developed to enable viewing both the raw transcripts and over 60 related text statistics together with their visualizations.

3. **comprehensive modeling.** The BERTopic topic modeling pipeline has been applied together with 5 different embedding methods and compared with the LDA baseline using topic coherence and diversity metrics, achieving superior results.

4. **performing thorough analysis.** The best BERTopic pipeline variant has been exploited to prepare a variety of topic modelling analyses, such as topic similarity, hierarchical dependencies and temporal changes.

Notably, while the UN General Debates corpus was previously available, its utilization by non-technical users was hindered by limited ease of access. Our developed interactive application serves as a bridge to overcome this constraint, offering a user-friendly platform that does not demand advanced programming skills. Beyond improved accessibility, the dataset has undergone enhancements which enable more thorough and nuanced analysis. We believe that the refined user interface empowers researchers, policymakers, and other non-technical users, enabling them to explore and extract valuable insights from the data more efficiently and comprehensively than before.

Future works. The next natural steps arise from our current efforts. Because of a lack of broad political science knowledge, we were only able to perform analyses at a basic, intuitive level. Therefore, we aim to prepare instructions for political scientists on how to use the interactive application. We also note that we limited ourselves to extracting only statistical features from the text. Applying methods that extract additional semantic features, such as sentiment analysis of polarity or subjectivity, will allow for gaining a more detailed insight into this complex dataset.

Ethical considerations

While machine learning algorithms may appear impartial, the inherent biases present in training data necessitate careful consideration of ethical implications, particularly in the realm of political science (Curini and Franzese, 2020). Acknowledging these challenges, we adhere to transparent data curation practices, ensuring the ethical deployment of NLP techniques in our research.

However, it is crucial to clarify and disclaim that this project does not constitute an official analysis of the UNGA corpus presented by any institution. Any insights derived from this study shall be employed solely for research purposes.

References

[Baturo et al.2017] Alexander Baturo, Niheer Dasandi, and Slava J. Mikhaylov. 2017. Understanding State Preferences with Text as Data: Introducing the UN General Debate Corpus. *Research & Politics*, 4(2):2053168017712821.

[Benoit et al.2018] Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. quanteda: An R Package for the Quantitative Analysis of Textual Data. *Journal of Open Source Software*, 3(30):774.

[Benzecri1992] Jean-Paul Benzecri. 1992. *Correspondence Analysis Handbook*. Textbooks and Monographs, New York.

[Blei and Lafferty2006] David M Blei and John D Lafferty. 2006. Dynamic Topic Models. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120.

[Blei et al.2003] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*.

[Calvo Bartolomé et al.2023] Lorena Calvo Bartolomé, José Antonio Espinosa Melchor, and Jerónimo Arenas-García. 2023. ITMT: Interactive Topic Model Trainer. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 43–49, Dubrovnik, Croatia, May. Association for Computational Linguistics.

[Chen et al.2023] Weisi Chen, Fethi Rabhi, Wenqi Liao, and Islam Al-Qudah. 2023. Leveraging State-of-the-Art Topic Modeling for News Impact Analysis on Financial Markets: A Comparative Study. *Electronics*, 12(12).

[Clapham et al.2022] Benjamin Clapham, Micha Bender, Jens Lausen, and Peter Gomber. 2022. Policy Making in the Financial Industry: A Framework for

Regulatory Impact Analysis Using Textual Analysis. *Journal of Business Economics*, pages 1–52.

[Curini and Franzese2020] Luigi Curini and Robert Franzese. 2020. *The SAGE Handbook of Research Methods in Political Science and International Relations*. SAGE.

[Dasandi et al.2023] Niheer Dasandi, Slava Jankin, and Alexander Baturo. 2023. Words to unite nations: The complete un general debate corpus, 1946-present, May.

[Devlin et al.2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

[Dieng et al.2020] Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

[Glavaš et al.2019] Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2019. Computational Analysis of Political Texts: Bridging Research Efforts Across Communities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. Association for Computational Linguistics.

[Greene and Cross2017] Derek Greene and James P Cross. 2017. Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. *Political Analysis*.

[Grimmer2010] Justin Grimmer. 2010. A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases. *Political Analysis*.

[Grootendorst2022] Maarten Grootendorst. 2022. BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure. *arXiv preprint arXiv:2203.05794*.

[Hansen et al.2023] Lasse Hansen, Ludvig Renbo Olsen, and Kenneth Enevoldsen. 2023. Textdescriptives: A python package for calculating a large variety of metrics from text. *Journal of Open Source Software*, 8(84):5153.

[Hollibaugh2018] Gary E Hollibaugh. 2018. The Use of Text as Data Methods in Public Administration: A Review and an Application to Agency Priorities. *Journal of Public Administration Research and Theory*.

[Honnibal et al.2020] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd, 2020. *spaCy: Industrial-strength Natural Language Processing in Python*.

[Jankin et al.2017] Slava Jankin, Alexander Baturo, and Niheer Dasandi, 2017. *United Nations General Debate Corpus 1946-2022*.

[Lau et al.2014] Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

[Laver et al.2003] Michael Laver, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2):311–331.

[Lee and Seung1999] Daniel D Lee and H Sebastian Seung. 1999. Learning the Parts of Objects by Non-negative Matrix Factorization. *Nature*.

[Lisena et al.2020] Pasquale Lisena, Ismail Harrando, Oussama Kandakji, and Raphael Troncy. 2020. TOMODAPI: A Topic Modeling API to Train, Use and Compare Topic Models. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 132–140, Online, November. Association for Computational Linguistics.

[Liu et al.2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

[Manning et al.2010] Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. 2010. Introduction to Information Retrieval. *Natural Language Engineering*, pages 100–103.

[McInnes et al.2017] Leland McInnes, John Healy, and Steve Astels. 2017. HDBSCAN: Hierarchical Density Based Clustering. *Journal of Open Source Software*, 2(11):205.

[McInnes et al.2018] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861.

[Nay2018] John Nay. 2018. Natural Language Processing and Machine Learning for Law and Policy Texts. *Available at SSRN 3438276*.

[Ng et al.2022] QX Ng, CE Yau, YL Lim, LKT Wong, and TM Liew. 2022. Public Sentiment on the Global Outbreak of Monkeypox: An Unsupervised Machine Learning Analysis of 352,182 Twitter Posts. *Public Health*, 213:1–4.

[Rehurek and Sojka2011] Radim Rehurek and Petr Sojka. 2011. Gensim–Python Framework for Vector Space Modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

[Reimers and Gurevych2019] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11.

[Röder et al.2015] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM International Conference on Web Search and Data Mining*, pages 399–408.

[Salton and Buckley1988] Gerard Salton and Christopher Buckley. 1988. Term-weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24(5):513–523.

[Sanh et al.2019] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *arXiv preprint arXiv:1910.01108*.

[Sievert and Shirley2014] Carson Sievert and Kenneth Shirley. 2014. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*.

[Terragni et al.2021] Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021. OCTIS: Comparing and Optimizing Topic models is Simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270, Online, April. Association for Computational Linguistics.

[Valdez et al.2018] Danny Valdez, Andrew C Pickett, and Patricia Goodson. 2018. Topic Modeling: Latent Semantic Analysis for the Social Sciences. *Social Science Quarterly*, 99(5):1665–1679.

[Vayansky and Kumar2020] Ike Vayansky and Sathish A.P. Kumar. 2020. A Review of Topic Modeling Methods. *Information Systems*.

[Wilczyński et al.2023] Piotr Wilczyński, Artur Żółkowski, Mateusz Krzyziński, Emilia Wiśnios, Bartosz Pieliński, Stanisław Giziński, Julian Sienkiewicz, and Przemysław Biecek. 2023. HADES: Homologous Automated Document Exploration and Summarization. *arXiv preprint arXiv:2302.13099*.

[Żółkowski et al.2022] Artur Żółkowski, Mateusz Krzyziński, Piotr Wilczyński, Stanisław Giziński, Emilia Wiśnios, Bartosz Pieliński, Julian Sienkiewicz, and Przemysław Biecek. 2022. Climate Policy Tracker: Pipeline for Automated Analysis of Public Climate Policies. *NeurIPS 2022 Workshop: Tackling Climate Change with Machine Learning*.

A Contribution

- Mateusz Grzyb – writing the proposal (introduction, rebuttal), preparing the presentations, performing the comparative model evaluation, describing the results of the experiment in the final report, preparing the reproducibility manual; **workload**: 8 hours per week;
- Mateusz Krzyziński – writing the proposal (related works, revision after reviews), preparing the presentations, preparing the interactive application, cleaning and enhancing the dataset, performing statistical analyses of the corpus, describing the process in the final report; **workload**: 8 hours per week;
- Bartłomiej Sobieski – writing the proposal (revision after reviews), preparing explanatory data analysis, preparing BERTopic analyses, describing them in the final report, preparing the presentations and examples; **workload**: 8 hours per week;
- Mikołaj Spytek – writing the proposal (abstract, approach & methodology, revision after reviews), preparing the presentations, performing the using different version of the BERTopic model, preparing a container of the final solution, describing experiments and results in the final report; **workload**: 8 hours per week;

B Application

This section contains Figures depicting visualizations available in the prepared interactive application.

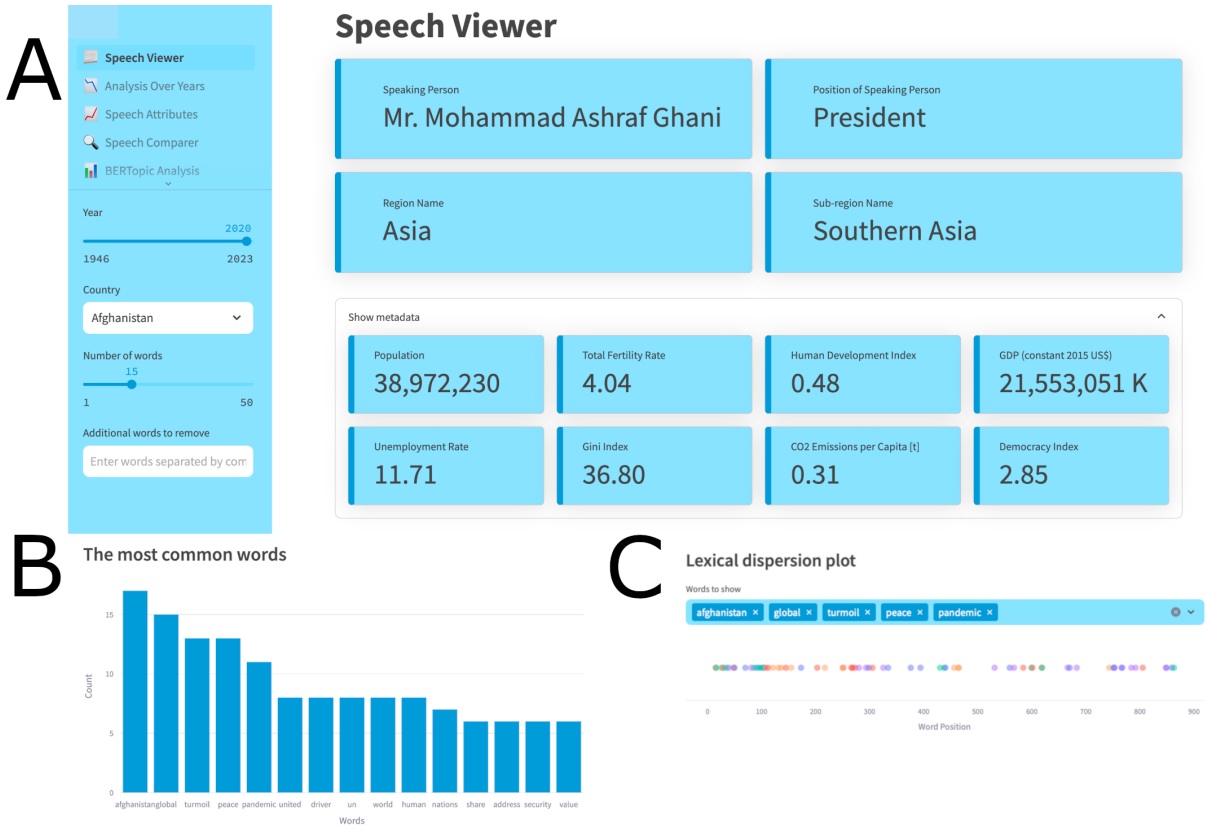

Figure 3: Speech Viewer: visualizations containing statistics about a selected speech. A – metadata about the selected speech; B – a histogram counting the occurrences of most common words; C – a lexical dispersion plot, showing at what positions words occur in the speech.

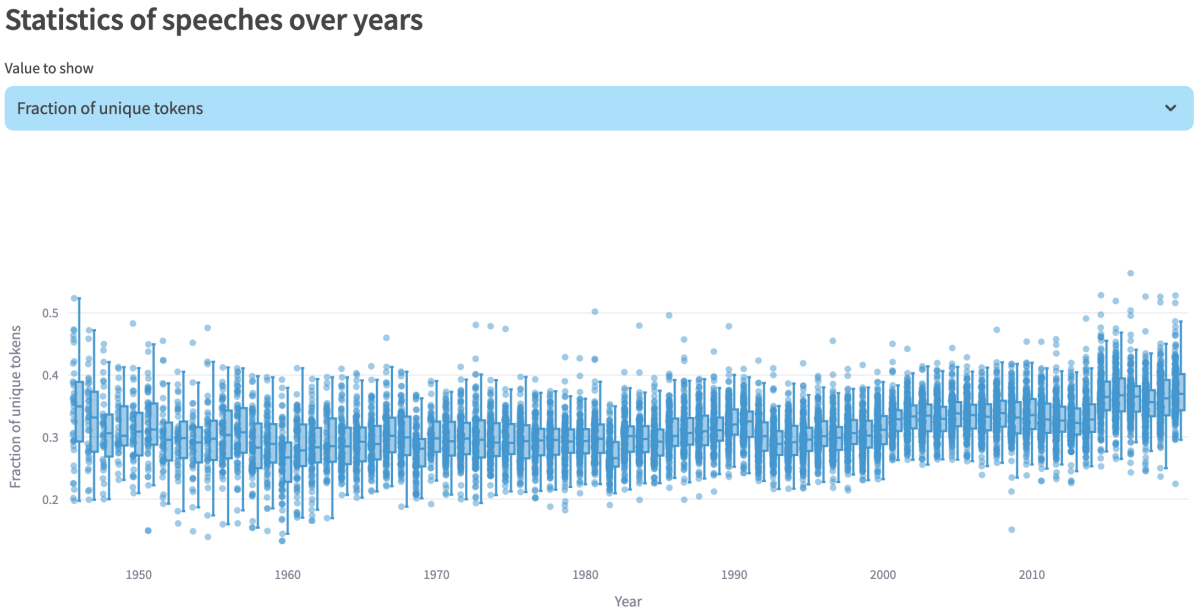

Figure 4: Example visualization from the Analysis Over Years panel depicting the distribution of the fraction of unique tokens in speeches from various countries over different years.

Figure 5: Example visualization from the Analysis Over Years panel depicting the relationship between the Flesch Readability Ease (readability index) and the democracy index for various countries spanning the years 2017 to 2022.

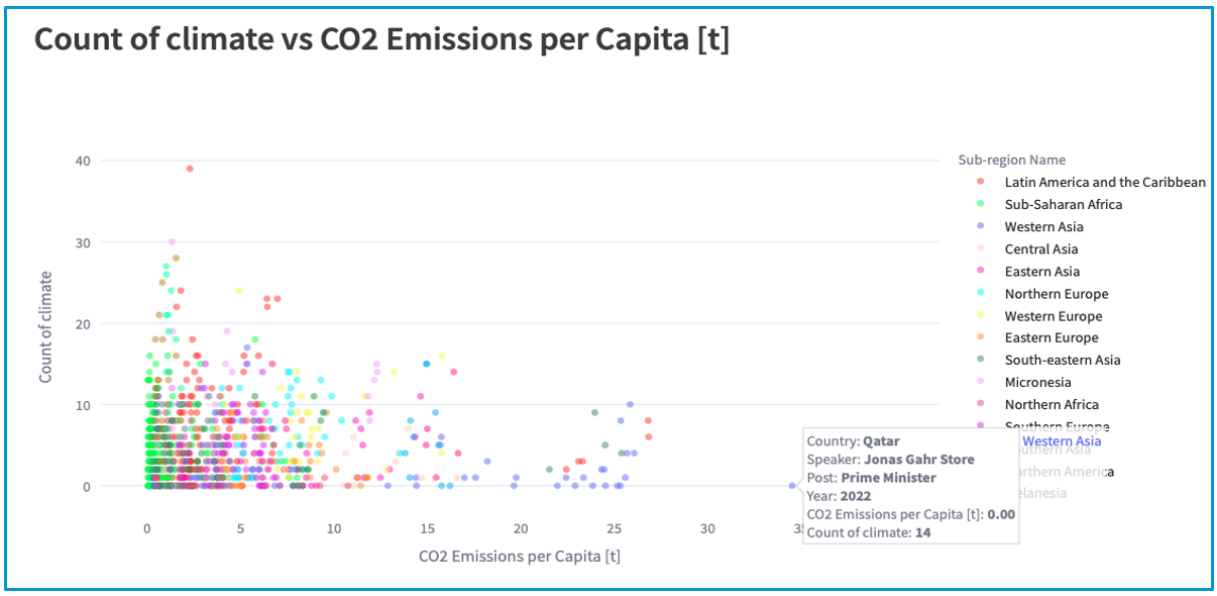

Figure 6: Example visualization from the Analysis Over Years panel depicting the frequency of the word 'climate' in speeches delivered by various countries from 2017 to 2022, juxtaposed against the corresponding CO2 emissions per capita.

C Reproducibility

- MODEL DESCRIPTION – Detailed description of the algorithm is contained in Section 3.4.
- LINK TO CODE – The source code is available in the subject repository at https://github.com/grant-TraDA/NLP-2023W/tree/main/13. Mining UNGA Debates/. Additionally the containerized application is available at https://hub.docker.com/r/krzyzinskim/nlp-unga-debates and hosted at https://nlp-unga-debates-g2o6gnzttq-lm.a.run.app/.
- INFRASTRUCTURE – All calculations were performed locally using personal laptops.
- RUNTIME PARAMETERS – Inference with the BERTopic model takes around 90 seconds. In the application, this action is only performed once. After that the results are cached.
- PARAMETERS – The used BERTopic model with `distillbert` embeddings has 110M parameters.
- VALIDATION PERFORMANCE – The performance of all models is reported in Table 3 and in Table 4. The comparison methodology is described in Section 4.2.
- METRICS – The used metrics are explained in Section 4.2. We use the implementation of the metrics from the OCTIS library (Terragni et al., 2021).

Multiple Experiments:

- NO TRAINING EVAL RUNS – *not applicable*, models were not trained from scratch. Pretrained models were used.
- HYPER BOUND – For the BERTopic model the following embedding models were tested: `all-MiniLM-L6-v2`, `all-MiniLM-L12-v2`, `all-mpnet-base-v2`, `distilbert` and `roberta`.
- HYPER BEST CONFIG – The `distilbert` embedding model performed the best.
- HYPER SEARCH – All possible embedding models were tested once.
- HYPER METHOD – All possible embedding models were tested, the selection criterion was the topic coherence metric.
- EXPECTED PERF – The summary statistics of the results are presented in Table 3 and in Table 4. The comparison methodology is described in Section 4.2.

Datasets – utilized in the experiments and/or the created ones:

- DATA STATS – The dataset contains 10 679 observations. One observation consists of one entire speech given at the UN General Assembly meeting.
- DATA SPLIT – The problem was unsupervised, so no data split was performed. The analysis was performed on 100% of the observations.
- DATA PROCESSING – Data preparation is described in Section 3.2.
- DATA DOWNLOAD – The original data is available at `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/0TJX8Y`. The data extended by us is available in the subject repository at https://github.com/grant-TraDA/NLP-2023W/tree/main/13. Mining UNGA Debates/.
- NEW DATA DESCRIPTION – The newly collected data is described in Section 3.1. The collected metadata with examples is presented with examples in Table 2.
- DATA LANGUAGES – All studied data was in English.