Mining United Nations General Assembly Debates
Project Proposal 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. Our aim is 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, our research project employs Natural Language Processing (NLP) techniques to extract meaningful insights from these transcripts, complemented by the collected metadata and statistical text analysis. 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 strategic alliances, uncovering region-specific policies, 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

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 project, we plan to 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 obtain answers to 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. (By the PoC stage – *22nd November*)
- Exploring the gathered corpus using statistical text analysis with the usage of collected metadata and visualisation of the results. (By the PoC stage – *22nd November*)
- 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. (By the final stage – *13th December*)

2 Related works

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 2.1. 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 de-

Table 1: Excerpt of the available metadata from (Jankin et al., 2017).

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

signed 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. 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.

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 comparison 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 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).

After tokenization and lemmatization, we exclude the stopwords, using the available, ready-made list for the English language. Moreover, we add bigrams and trigrams to individual words to improve the topic modeling and streamline topic interpretation.

3.2 Statistical textual data analysis

Simple statistical analysis of the text corpora can be used to extract insightful information without sophisticated modeling methods. Frameworks, such as quanteda proposed by Benoit et al. (2018) facilitate these analysis.

These frameworks offer many tools, ranging from simple frequency analysis methods to see how the distribution of tokens changes in different partitions of the corpus, through coalition analysis to study set phrases in which words occur together, to the analysis of lexical diversity in particular documents and creating hierarchical structures of documents based on their similarity.

For mining General Assembly debates these tools are particularly helpful when considering the evolution of changes in the speeches of particular countries over time, as well as comparing the features of documents coming from different regions.

3.3 BERTopic

Theoretical setup. The BERTopic model uses a pre-trained 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.

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 TF-IDF 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 rel-

ative 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.

3.4 Technical considerations

The majority of the project uses the Python programming language. The required versions of libraries needed to reproduce results of this project are listed in the `requirements.txt` file attached to the source code.

When necessary, the research for this project will be carried out with the computational resources made available by Google Colab. In the cases where the free resources from this platform are not sufficient parts of the project may be carried out thanks to the support of the Laboratory of Bioinformatics and Computational Genomics and the High Performance Computing Center of the Faculty of Mathematics and Information Science Warsaw University of Technology on the High Performance Computing cluster.

References

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.

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.

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

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

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

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.

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).

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.

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

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.

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.

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

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

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

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*.

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

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

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.

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

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.

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

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

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

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

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.

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).

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

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.

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.

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

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*.

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*.