Topic Analysis of UN Speeches (1971 to 2018)

Using Latent Dirichlet Allocation and Topic GPT

Ritesh Kumar

2024SP_MS_DSP_453-DL_SEC61: Natural Language Processing

Project P.4

Nethra Sambamoorthi and Sudha BG

June 10, 2024

Table of Contents

Introduction

The United Nations General Assembly (UNGA) stands as a pivotal platform where representatives from over 190 nations convene to address and deliberate on the world's most pressing issues. Since its inception, the UNGA has provided a stage for member states to articulate their national interests, express concerns, and outline their visions for the future. The speeches delivered during these sessions offer a rich tapestry of global perspectives, reflecting the evolving priorities and challenges faced by the international community. By analyzing the topics discussed in UNGA speeches from 1971 to 2018, we can gain unique insights into the highlights of international dynamics, the discussions of global challenges, and the themes of diplomatic focuses over nearly five decades.

This project leverages advanced topic modeling techniques, specifically Latent Dirichlet Allocation (LDA) and Topic GPT, to uncover the thematic structures embedded within these speeches. LDA is a probabilistic topic modeling technique that helps identify latent topics in large collections of text. Topic GPT, on the other hand, uses contextual embeddings and clustering to discover more nuanced and coherent topics. By applying these models to UNGA speeches, we aim to systematically identify the dominant topics discussed by member states over time.

Problem Statement

Understanding the topics discussed in UNGA speeches is crucial for several reasons. First, it provides insights into the historical context of international relations and highlights of global discussions. Second, it helps identify emerging challenges and trends that have shaped the global agenda. Third, it enhances our understanding of the diplomatic focuses of different countries, offering a comparative perspective on national interests and concerns. Despite the significance of these speeches, there has been limited systematic analysis of their content

over an extended period. This project addresses this gap by applying advanced topic modeling techniques to a comprehensive dataset of UNGA speeches, aiming to reveal the underlying topics and their development over nearly five decades.

The specific objectives of this project are as follows:

1. Data Collection and Organization: Compile a comprehensive archive of UNGA speeches from 1971 to 2018, ensuring the data is well-structured and organized for analysis. This involves extracting and reading text files, capturing metadata, and compiling the data into a format suitable for topic modeling.
2. Topic Modeling with LDA and Topic GPT:
   - LDA: Apply LDA to the dataset to identify the dominant topics discussed in the speeches. This involves preprocessing the text data to remove noise, fitting the LDA model, and interpreting the resulting topics.
   - Topic GPT: Use Topic GPT to leverage contextual embeddings and clustering for a more sophisticated topic modeling approach. This technique can capture the nuances and variations in speech content, providing a richer analysis of the thematic structures.
3. Identification of Dominant Topics: Systematically identify and catalog the main topics discussed in the UNGA speeches over the years.
4. Comparative Analysis by Country: Examine the topics emphasized by different countries to uncover patterns in national interests and concerns. This involves comparing the topics discussed by various countries and exploring how their diplomatic priorities have evolved.
5. Visualization and Reporting: Develop visualizations to effectively communicate the findings of the analysis.

By employing these advanced topic modeling techniques, we aim to provide a comprehensive and nuanced understanding of the themes and issues that have shaped UNGA discussions over nearly five decades.

Literature Review

Topic modeling is a fundamental task in natural language processing (NLP) aimed at discovering the underlying thematic structures in text corpora. Traditional models such as Latent Dirichlet Allocation (LDA) and more contemporary approaches like GPT-based topic modeling have significantly advanced this field. This literature review provides an overview of these methodologies, emphasizing their contributions, limitations, and comparative performance.

Latent Dirichlet Allocation (LDA), introduced by Blei, Ng, and Jordan in 2003, is a generative probabilistic model designed to uncover hidden topics within a corpus of documents. Each document is represented as a mixture of topics, where each topic is characterized by a distribution over words. LDA assumes a Dirichlet prior distribution over the topic distributions of documents and the word distributions of topics, enabling it to effectively capture the probabilistic relationships between words and topics.

One of the key strengths of LDA is its ability to provide interpretable topics through word distributions, making it a valuable tool for various text mining applications, including document classification, similarity measurement, and text summarization. However, LDA's reliance on the bag-of-words representation neglects the semantic relationships between words, potentially limiting its ability to capture contextual nuances.

GPT and Neural Topic Models**:** The advent of Generative Pre-trained Transformers (GPT) has revolutionized NLP by providing rich, contextual embeddings of text. GPT, developed by

OpenAI, pre-trains deep bidirectional representations by conditioning on large amounts of text data, resulting in superior performance on a wide range of NLP tasks.

Topic GPT, a novel approach leveraging GPT embeddings for topic modeling, integrates pre-trained transformer-based language models to generate document embeddings, which are then clustered to form topics. This method addresses the limitations of traditional models like LDA by capturing semantic relationships through contextual embeddings. A notable paper on this approach is "Topic GPT: A Prompt-Based Topic Modelling Framework" by Pham et al. (2024), which discusses the methodology and performance of Topic GPT in detail.

Topic GPT utilizes a class-based variation of TF-IDF to extract topic representations, ensuring coherent and distinct topic formation. This approach enhances the interpretability and relevance of the generated topics compared to those produced by LDA. Additionally, Topic GPT's flexibility in incorporating state-of-the-art language models allows it to remain competitive across various benchmarks.

Comparative Performance and Applications: Empirical evaluations indicate that Topic GPT outperforms LDA in terms of topic coherence and diversity, especially in datasets where semantic context plays a crucial role. For instance, experiments involving the 20 NewsGroups and BBC News datasets demonstrate Topic GPT's superiority in generating coherent and semantically rich topics.

However, LDA remains a robust choice for large-scale text corpora where computational efficiency is paramount. Its simplicity and well-established theoretical foundation make it suitable for applications in information retrieval and document classification.

Conclusion: The evolution from LDA to GPT-based topic models like Topic GPT represents a significant leap in the field of topic modeling. While LDA provides a foundational framework for probabilistic topic discovery, the integration of neural embeddings in Topic GPT offers

enhanced semantic understanding and topic coherence. Future research may focus on further refining these models, exploring hybrid approaches, and extending their applicability to diverse NLP tasks.

Data

Source and Structure of Data: The data for this project consists of a comprehensive archive of United Nations General Assembly speeches spanning from 1971 to 2018. Originally stored in a zipped folder of 68.6MB, the archive encompasses a structured collection of text files organized by session and year. Each session folder is named systematically, such as "Session 25 - 1970", and contains multiple text files corresponding to speeches delivered by different countries, formatted as "COUNTRYCODE_YEAR.txt" (e.g., "USA_1970.txt").

Unzipping and Organizing: The extraction process began with the decompression of the zipped archive. Using Python's `zipfile` module, the compressed data was programmatically extracted into a designated directory, maintaining the original hierarchical structure of session folders.

Data Reading and Preliminary Processing: Each session folder was iterated overusing Python's `os` and `glob` modules to navigate folders and access files. Text files within these folders were read sequentially, with the speech text being loaded into memory. Simultaneously, metadata extracted from the file names and directory structure—specifically, the country code, session number, and year—was recorded. This allowed for each speech to be associated with its contextual details, crucial for subsequent analytical stages.

Compilation into Analytical Format: After extracting the text and metadata, the speeches were compiled into a structured format suitable for processing with LDA. Using the `pandas` DataFrame, each speech was represented as a row, with columns for country, year, session, and the raw text of the speech. This structured format facilitated efficient manipulation and transformation of the data necessary for topic modeling.

Preprocessing

Removal of Numeric Identifiers: Speeches often contained paragraph numbers or other numeric identifiers that are irrelevant to text analysis. These were removed to avoid misleading the topic modeling process, which should focus solely on substantive textual content.

Text Normalization:

- New Line Characters: To ensure the text was processed as continuous prose, new line characters were replaced with spaces. This helps in maintaining the natural flow of sentences and paragraphs without arbitrary breaks introduced by formatting.
- Apostrophes and Quotation Marks: These were removed to standardize the text, reducing variations of words and phrases (e.g., transforming contractions to their full forms).
- Hyphens: Often used in compound words or hyphenated names, these were removed to avoid splitting terms that are typically analyzed as a single unit in natural language processing.

Standardization of Whitespace: Excess whitespace, including spaces resulting from the removal of punctuation or special characters, was standardized. This involved stripping leading and trailing spaces and reducing instances of multiple spaces between words to a single space. This step is crucial for maintaining consistent tokenization during the text analysis phase.

Removal of Salutations: Common salutations such as "Mr.", "Mrs.", and "Dr." were removed. These elements are generally not informative for the analysis of topics and removing them helps to focus the topic modeling on more meaningful content.

Handling Special Text Elements: References to external texts or annotations within the speeches, which could distract from the main content, were removed. This includes any

textual content within square brackets or similar notations that are not part of the primary speech content.

Sentence Segmentation: After cleaning, the text was segmented into individual sentences. This step is essential for certain types of analysis where the context of a sentence is necessary to understand the usage and meaning of words and phrases within the specific grammatical structure.

These preprocessing steps ensured that the text data was clean, uniform, and well-suited for extracting meaningful insights through topic modeling. By standardizing the format and removing irrelevant elements, the cleaned data provides a more accurate foundation for identifying the key themes and topics discussed in the UN General Assembly over the years.

After the textual data from the UN speeches was preprocessed, Latent Dirichlet Allocation (LDA) and TopicGPT were employed to analyze and uncover the thematic structures within these texts.

To understand the variation in speech lengths and participation frequency of countries, we created boxplots representing the average number of words per speech and the average number of speeches made by each country.

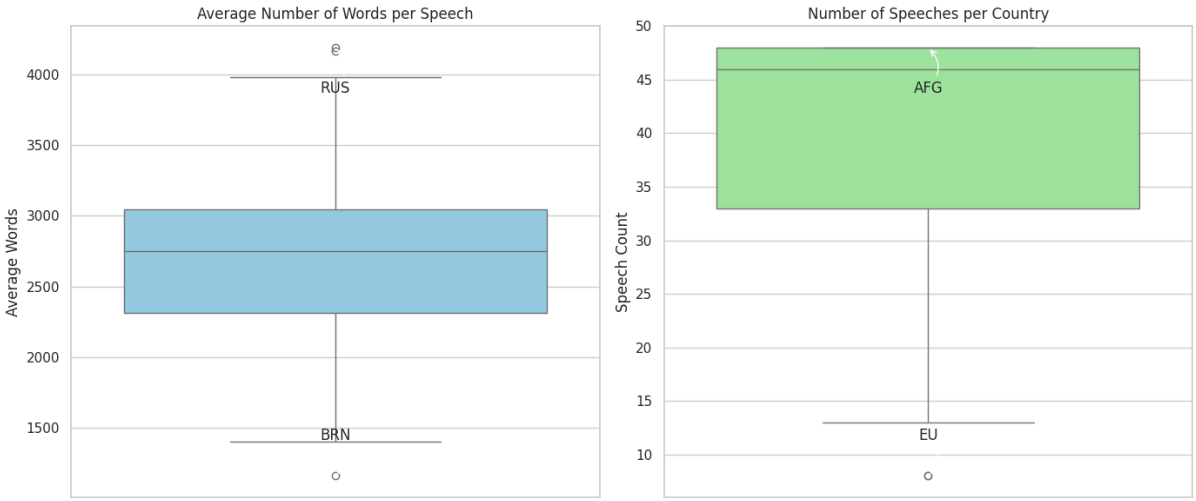

Figure 1

The boxplots provide a comparative analysis of UNGA speech data from 1971 to 2018. The first plot illustrates that Russia (RUS) consistently delivered longer speeches on average, with significantly higher word counts compared to other nations. The second plot highlights Afghanistan (AFG) as the country with the highest total number of speeches delivered at the UNGA during the observed period, indicating a very active participation. This contrasts with other countries like the European Union (EU), which had a notably lower count of speeches, suggesting less frequent participation or representation in terms of speech quantity at these sessions.

Research and Modeling Methods

In this project, we employed Latent Dirichlet Allocation (LDA) and TopicGPT to analyze United Nations General Assembly (UNGA) speeches from 1971 to 2018, revealing the thematic structures within these texts.

Latent Dirichlet Allocation (LDA) is a generative probabilistic model designed to discover latent topics in large collections of text. This method represents each document as a mixture of topics, with each topic characterized by a distribution over words. LDA's generative process involves assuming a Dirichlet prior distribution over the topic distributions for each document and the word distributions for each topic. This approach effectively captures the probabilistic relationships between words and topics, making it suitable for uncovering hidden themes in text corpora.

The preprocessing steps for LDA include several crucial tasks:

1. Noise Removal: Removing irrelevant data, such as punctuation, numbers, and special characters, which do not contribute to the meaning of the text.
2. Text Normalization: Converting all text to a uniform format, such as lowercasing all words to ensure consistency.

3. Data Structuring: Organizing the text data into a suitable format, typically by tokenizing the text into words or phrases and representing documents as bags-of-words. This ensures that the model focuses solely on substantive textual content.

TopicGPT leverages the power of Generative Pre-trained Transformers (GPT) for topic modeling. Unlike traditional models like LDA, TopicGPT uses contextual embeddings to generate document representations, which are then clustered to form topics. This method addresses the limitations of LDA by capturing semantic relationships through GPT embeddings. The model generates document embeddings using a pre-trained transformer-based language model, clusters these embeddings, and applies a class-based variation of TF-IDF to extract coherent and distinct topic representations.

The preprocessing steps for TopicGPT are as follows:

1. Text Normalization: Similar to LDA, this step ensures the text is in a uniform format, facilitating better processing by the model.

2. Handling Special Elements: Removing special text elements such as salutations and annotations that do not contribute to the main content.

3. Data Cleaning: Ensuring the data is free from noise and irrelevant elements, making it suitable for analysis.

4. Sentence Segmentation: Segmenting text into individual sentences to maintain the context and grammatical structure.

TopicGPT's ability to leverage state-of-the-art language models ensures it remains competitive across various benchmarks. This model enhances interpretability and relevance by applying a sophisticated approach using contextual embeddings to capture nuances in speech content.

Application of LDA and TopicGPT to UNGA Speeches: By applying these models to the comprehensive dataset of UNGA speeches, we systematically identify the dominant topics discussed by member states over time. LDA provides a probabilistic framework for discovering hidden topics, while TopicGPT offers a more nuanced approach using contextual embeddings. Together, these models enable a deeper understanding of the international themes and diplomatic focuses within the UNGA speeches over nearly five decades. This analysis provides insights into the historical context of international relations, emerging global challenges, and the diplomatic priorities of different countries.

Data Processing

A script was designed to extract, preprocess, and analyze text data from the speeches stored in text files.

Overview of the Functionality

1. Defined the Base Path: The `base_path` variable specified the root directory ("Converted sessions/") where session folders were stored, each corresponding to a different year of UN General Assembly sessions.
2. Iterated Over Session Folders: The script used `glob.glob(base_path + 'Session')` to find all directories that matched the pattern, which were session folders labeled by year (e.g., "Session 25 - 1970"). It then iterated through each session folder, extracted Session and Year: The session name and year were extracted from the folder name using string splitting operations. This metadata was crucial for tracking the source of each speech.
3. Iterated Over Files in Each Session Folder: For each session, the script iterated over all `.txt` files (representing individual country speeches for that session year). It used `glob.glob(session_folder + '/.txt')` to list all text files in the current session folder. The country code was extracted from each file name, which helped in identifying the country each speech represented.

4. Read and Preprocessed Speech Text:
    - File Reading: Each file was opened and read, loading the text of the speech into memory.
    - Cleaning and Sentence Splitting: The `clean` function (assumed to exist) was applied to preprocess the text by removing unwanted characters, standardizing formatting, etc. After cleaning, the `sentences` function (also assumed to exist) was used to split the cleaned text into individual sentences, creating a structured representation of the speech.
5. Combined Sentences and Extracted Topics:
    - Combined Sentences: If the result from `sentences` was a list (as expected after splitting, ensuring it was in the correct format for topic modeling.
    - Topic Extraction: The `extract_topics_single` function was applied to the combined speech to identify and extract the main topics using LDA or a similar method. This function presumably transformed the text into a format suitable for LDA, ran the analysis, and returned the identified topics.
6. Stored Results: Each speech's data—including the country code, session, year, the combined (processed) speech text, and the extracted topics—was stored as a dictionary in the `data` list. This structured format was beneficial for subsequent analysis and review.
7. Created DataFrame: Finally, all collected data dictionaries were converted into a Pandas DataFrame. This DataFrame was an efficient data structure for handling large datasets and supported various operations necessary for data analysis, visualization, and reporting.

This script effectively automated the process of extracting, cleaning, and analyzing a large volume of textual data from structured directories. It converted raw text files into a structured dataset ready for advanced textual analysis.

	Country	Session	Year	Speech	Topics
0	PER	Session 26 - 1971	1971	Mr President, I am very pleased to be able now...	[world, international, countries, new, nations]
1	FJI	Session 26 - 1971	1971	181 Mr President, may I, on behalf of my dele...	[fiji, nations, pacific, small, united]
2	SDN	Session 26 - 1971	1971	Your election to this office, Mi President, i...	[world, nations, united, united nations, mr]
3	ETH	Session 26 - 1971	1971	Mr President, it is my very pleasant duty to e...	[nations, united, united nations, general, int...
4	MAR	Session 26 - 1971	1971	Mr President, first of all, I should like to a...	[united, international, nations, united nation...

Figure 2

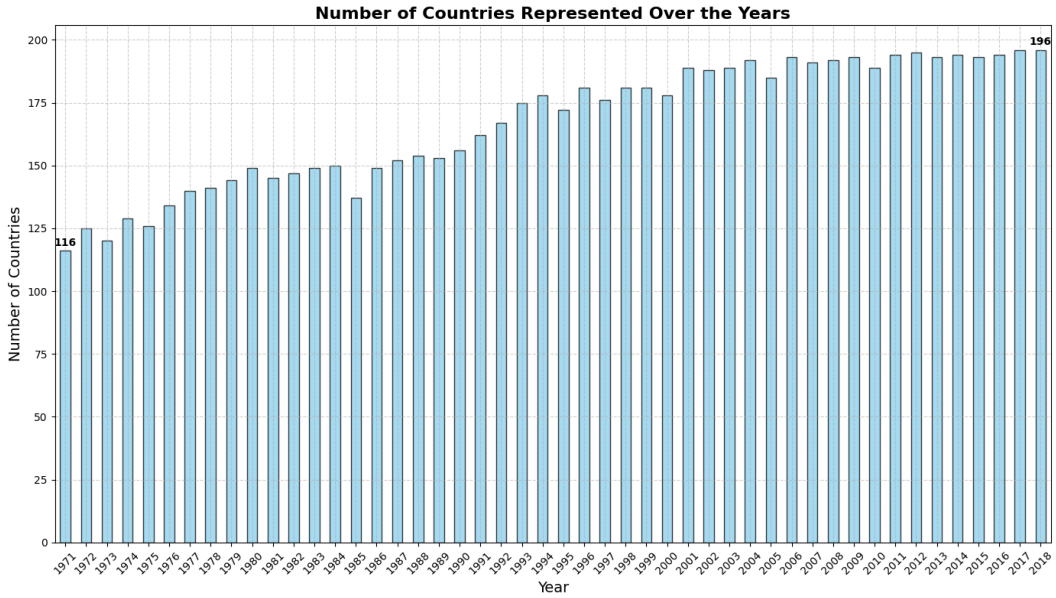

Figure 3

The bar chart displays a steady increase in the number of countries represented at the UN.

```
[154] df.Speech[67]

'Mr President, Egypt, which has with your country and your friendly people the closest ties of history and culture and £
a common struggle, is truly happy to see you assume your high post as President of the General Assembly  It is also a gre
at pleasure for me to congratulate you on the assumption of your high office  With the outstanding qualities that you pos
sess you will indeed guide this session to important achievements You succeed in this high office Edvard Hambro of Norwa
y, whose name will remain linked with the historic declarations adopted last year under his able leadership  These declar
ations will continue to be of great value to the United Nations order and to co operation among States Ten years ago the
United Nations lost Dag Hammarskjold  Today, after 10 eventful years through which U Thant has guided the United Nations
with unique ability and integrity, all who know the Secretary General should rejoice in the success he has achieved and t
he values with which he has enrich...'

df.Topics[67]

['israel, security, peace, resolution, united']
```

Figure 4: Sample Speech and Extracted Topics

Extracting topics using Latent Dirichlet Allocation (LDA) was a relatively straightforward process that completed in just a few minutes. The simplicity of LDA, coupled with its

probabilistic framework, allows for quick identification and categorization of latent topics within large text corpora. This efficiency is one of the key advantages of using LDA for topic modeling, especially when dealing with large datasets. Moreover, LDA is a cost-effective solution as it does not require any external API keys or associated expenses, making it freely accessible for academic and research purposes.

On the other hand, extracting topics using TopicGPT was a considerably more complex and time-intensive process, requiring several hours to complete. The advanced nature of TopicGPT, which leverages the power of Generative Pre-trained Transformers (GPT), contributes to its longer processing time. TopicGPT generates rich, contextual embeddings for each document, capturing intricate semantic relationships that traditional models like LDA might miss. However, this complexity comes with the trade-off of increased computational demands and processing time.

Additionally, using TopicGPT necessitated obtaining an API key and incurred an expense of approximately $25. The requirement of an API key and associated costs add a financial consideration that was not present with LDA. This financial investment, although relatively modest, is an important factor to consider, especially for extensive or large-scale analyses.

Moreover, the output generated by TopicGPT was not in a uniform format. Often, a single entry contained multiple topics and associated keywords and that too in multiple formats, which made the results less straightforward to interpret and compare directly with those from LDA. This lack of uniformity in TopicGPT's output required additional steps in the data processing pipeline.

To facilitate a meaningful comparison between the outputs from LDA and TopicGPT, we undertook a rigorous data cleaning process. This involved parsing through the TopicGPT results to extract and standardize the relevant information. Specifically, we created a list of

five words/phrases for each entry, consisting of the first identified topic followed by four related keywords. This step was crucial in ensuring that the outputs from both models could be compared on a like-for-like basis, thereby enabling a more accurate and insightful analysis of the thematic structures within the UNGA speeches.

In summary, while LDA offers the advantage of quick and efficient topic extraction at no cost, TopicGPT provides a deeper, more nuanced understanding of the text at the expense of increased processing time, complexity, and cost. The additional data cleaning required for TopicGPT highlights the importance of post-processing in extracting valuable insights from advanced NLP models. This meticulous approach allowed us to harness the strengths of both LDA and TopicGPT.

Results

These lists show the top 25 most frequent topics extracted using Latent Dirichlet Allocation (LDA) and Topic GPT from the UNGA speeches dataset. Here is a comparative analysis and interpretation:

LDA: Top 25 Most Frequent Topics:

nations: 4562
united: 4045
international: 3528
world: 2487
united nations: 2435
countries: 2042
peace: 1732
development: 1317
people: 1062
security: 820
states: 819
economic: 548
country: 471
africa: 426
government: 359
new: 358
human: 339
general: 317
rights: 287
republic: 259
global: 225
nuclear: 201
south: 185
organization: 179
council: 172

Topic GPT: Top 25 Most Frequent Topics:

United Nations: 3496
General Assembly: 2110
International Relations: 878
Security Council: 863
Secretary General: 795
Terrorism: 532
Peace: 463
United Nations and International Relations: 431
United Nations Reform: 423
peace: 402
International cooperation: 392
Millennium Development Goals: 322
Human rights: 321
President: 309
Peacekeeping: 304
Conflict resolution: 303
Disarmament: 296
Multilateralism: 243
international community: 235
Climate change: 226
International Relations and Diplomacy: 216
Peace and security: 209
Sustainable development: 194
Diplomacy: 192
cooperation: 187

Figure 5: Top 25 Most Frequent Topics

LDA: Top 25 Most Frequent Topics

LDA's output reveals a high frequency of terms such as "nations" (4562), "united" (4045), "international" (3528), and "world" (2487). These terms indicate the centrality of global cooperation and international relations in UNGA speeches. Additionally, terms like "united nations" (2435) and "countries" (2042) frequently appear, reflecting discussions on global governance and the roles of member states.

Key issues such as "peace" (1732), "development" (1317), and "security" (820) underscore their importance in international discourse. Economic and social concerns are also prominent, with terms like "economic" (548), "people" (1062), and "government" (359) indicating a focus on human development and governance. The emphasis on specific regions, such as Africa ("africa" (426)) and states ("states" (819)), points to targeted discussions on regional challenges and development. Human rights ("rights" (287)), global issues ("global" (225)), and nuclear disarmament ("nuclear" (201)) are significant topics, reflecting the UN's broader mission.

Topic GPT: Top 25 Most Frequent Topics

Topic GPT's output highlights terms like "United Nations" (3496) and "General Assembly" (2110), emphasizing the centrality of the UN and its primary body in global discussions. Governance and security-related terms, such as "Security Council" (863) and "Secretary General" (795), are prominent, reflecting the institutional focus of the UNGA.

Peace and security issues, including "Terrorism" (532), "Peace" (463), and "Peacekeeping" (304), appear frequently, indicating their critical importance. The presence of terms like "United Nations and International Relations" (431) and "United Nations Reform" (423) highlights discussions on governance reforms and international cooperation.

Development-related terms, such as "Millennium Development Goals" (322) and "Sustainable development" (194), suggest a significant focus on global development agendas. The inclusion of terms like "Human rights" (321), "Climate change" (226), and "Disarmament" (296) indicates a comprehensive approach to addressing various global challenges.

Comparative Insights

Both LDA and Topic GPT outputs highlight the significance of the United Nations, international cooperation, and peace in the UNGA speeches. However, Topic GPT provides a more nuanced view by capturing specific entities and more complex themes, such as "Millennium Development Goals" and "Climate change," which are not explicitly identified in LDA's output. This indicates Topic GPT's ability to capture richer contextual information and semantic relationships in the text.

LDA tends to focus on more general terms and broader themes, while Topic GPT delves deeper into specific topics and provides a more detailed thematic structure. The need for additional data cleaning and processing for Topic GPT's output underscores its complexity but also its potential for yielding more insightful and granular results.

In conclusion, while both methods highlight critical themes in UNGA speeches, Topic GPT offers a more detailed and nuanced understanding of the text, capturing specific topics and their relationships more effectively than LDA. This comprehensive analysis provides valuable insights into the historical context of international relations, emerging global challenges, and the diplomatic priorities of different countries over nearly five decades.

Analysis and Interpretation

To compare the outputs from LDA and TopicGPT, we enhanced our analysis by creating word clouds that visualize the top 100 topics identified by each model. This approach

allowed us to visually assess the thematic differences and similarities between the two methods, providing a clearer understanding of the textual data's underlying structure.

LDA: Top 100 Topics Word Cloud

Topic GPT: Top 100 Topics Word Cloud

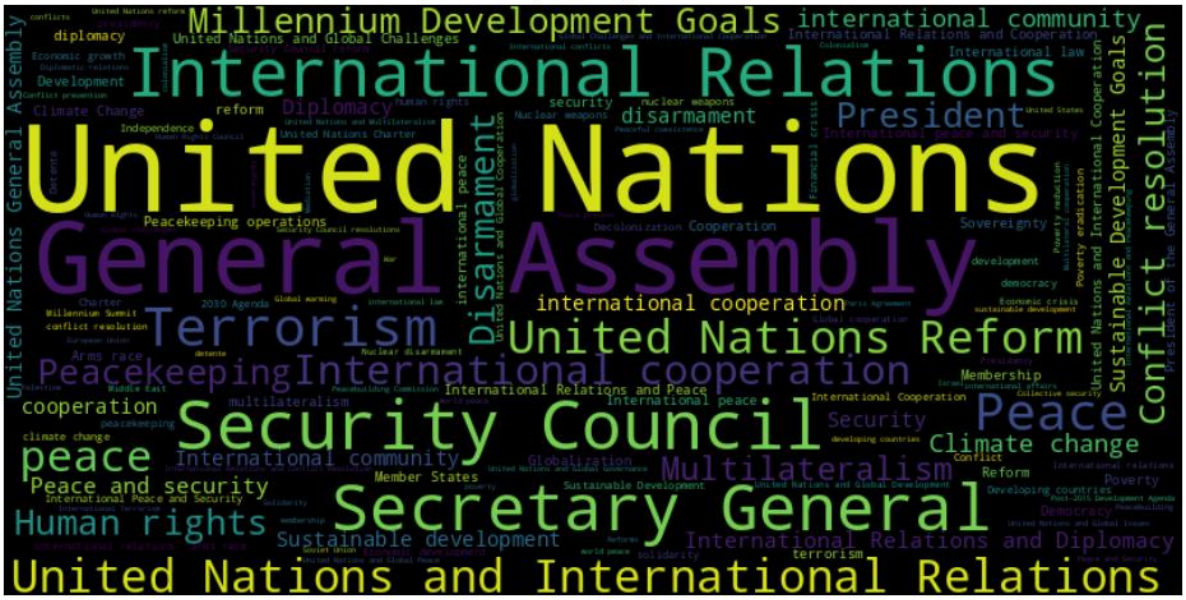

Figure 6: Word Clouds

The word clouds generated from the top 100 topics identified by LDA and Topic GPT offer a visual comparison of the thematic structures within UNGA speeches, revealing distinct insights into the discussions at the United Nations General Assembly.

The Topic GPT word cloud prominently features terms like "United Nations," "General Assembly," and "International Relations," reflecting the central focus on global governance and diplomatic interactions. It also highlights specific entities and initiatives such as the "Security Council," "Secretary General," and "Peacekeeping," underscoring the emphasis on governance and peace operations. Complex themes like "Millennium Development Goals," "Climate Change," and "Sustainable Development" indicate the Assembly's engagement with contemporary global challenges. Additionally, terms like "Disarmament," "Terrorism," and "Conflict Resolution" illustrate the critical importance of security issues. The frequent mention of "United Nations Reform" and "International Cooperation" further emphasizes the UN's continuous efforts to improve and collaborate globally.

In contrast, the LDA word cloud focuses on broader themes and general terms. Dominant terms such as "nations," "united," "international," and "world" underscore the global cooperation discussed at the Assembly. Key issues like "peace," "development," and "security" highlight the importance of these topics in the speeches. The LDA word cloud also reveals significant attention to economic and social concerns with terms like "economic," "people," and "government," indicating a focus on human development and governance. Regional mentions, including "Africa" and "states," suggest targeted discussions on regional challenges and development. Terms related to human rights, such as "rights," and global issues like "nuclear," reflect the UN's broader mission.

Comparing these visualizations, Topic GPT captures more specific and nuanced themes, reflecting complex issues and initiatives, while LDA provides a broad overview of general themes and issues. Topic GPT's ability to leverage contextual embeddings results in richer, more detailed thematic structures, whereas LDA, relying on the bag-of-words model, misses some of the nuanced relationships between words and themes.

In conclusion, while LDA offers a quick and broad overview of topics, Topic GPT provides a deeper, more nuanced analysis, making it a powerful tool for understanding the complex textual data of UNGA speeches. These insights enable a comprehensive analysis of the historical context of international relations, emerging global challenges, and the diplomatic priorities of different countries over nearly five decades.

Clustering

To comprehensively analyze the United Nations General Assembly (UNGA) speeches, we applied clustering techniques to topics derived from both Latent Dirichlet Allocation (LDA) and Topic GPT. This dual approach allowed us to capture a broader and more detailed thematic landscape.

We began by converting the topics generated by Topic GPT into TF-IDF vectors to quantify the importance of each term within the dataset, enabling us to handle the textual data numerically. We then explored a range of parameters to identify the optimal configuration for the KMeans clustering algorithm. This involved varying the number of clusters, initialization methods, and the number of initializations. Each combination was evaluated using the silhouette score, which measures how similar an object is to its own cluster compared to other clusters. After extensive testing, the optimal configuration for Topic GPT was found to be two clusters, with the `k-means++` initialization method and ten initializations. This configuration achieved a silhouette score of 0.2139, indicating distinct and meaningful clusters, though with some room for improvement in cohesion.

Similarly, we transformed the topics derived from LDA into TF-IDF vectors, ensuring consistency in comparison between the two models. We employed the same rigorous parameter tuning process, testing various combinations to identify the best clustering configuration. For LDA, the best setup was also two clusters but required twenty initializations using the `k-means++` method. This achieved a silhouette score of 0.2896, indicating higher cohesion within clusters and better separation between clusters compared to the Topic GPT configuration.

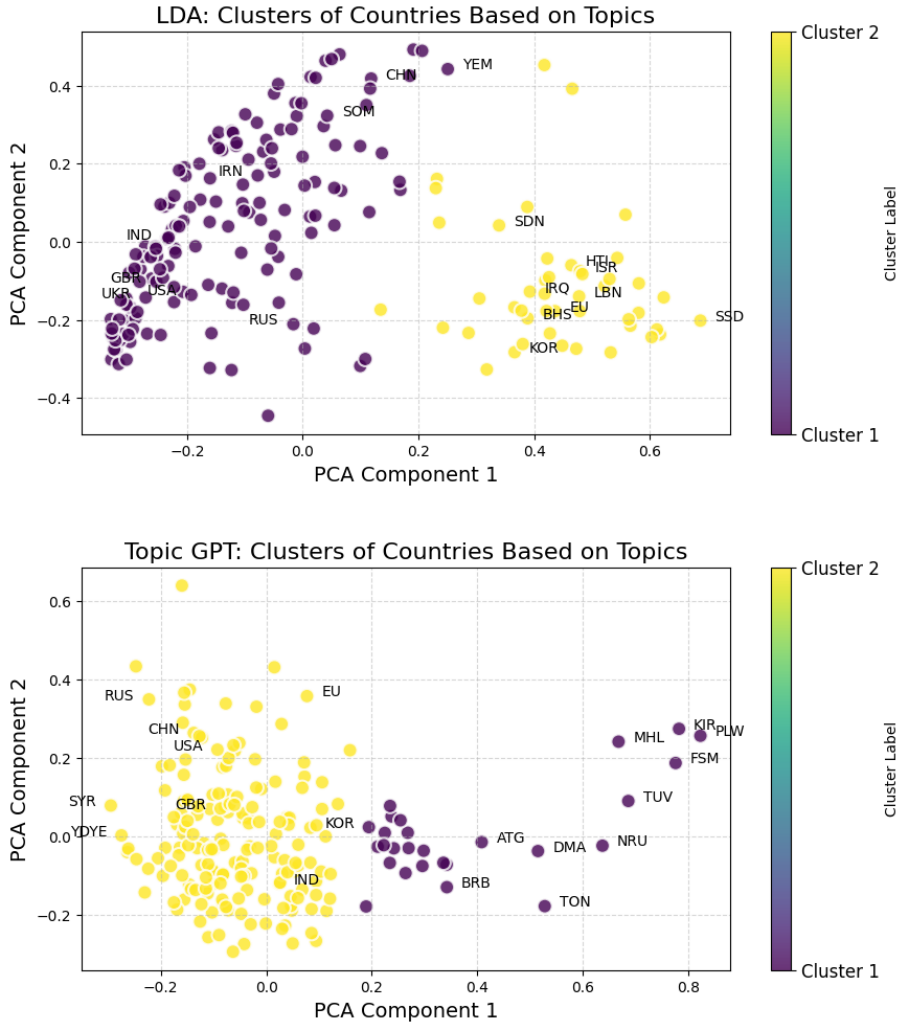

Figure 7: Clusters

The LDA plot reveals two clusters: Cluster 2 (purple) includes major global players like the USA, UK, Russia, and China, indicating common themes and strategic issues. Cluster 1

(yellow) includes countries like Iraq, South Sudan, and Haiti, focusing on regional or developmental issues. The first principal component likely captures a gradient from global governance to regional topics, while the second component might represent economic versus security issues.

The Topic GPT plot also shows two clusters but with different alignments. Cluster 1 (yellow) includes global players such as Russia, China, and the USA, indicating a focus on international relations and governance. Cluster 2 (purple) contains smaller countries like Tonga and Nauru, emphasizing local or regional issues. The first principal component likely represents a spectrum from global to regional focus, and the second component may indicate economic versus security concerns.

Comparative Analysis: Both models identify two clusters but group countries differently, reflecting their distinct thematic extractions. LDA provides a broad overview of major themes, while Topic GPT offers nuanced insights into specific entities and relationships. LDA's clustering shows a mixed distribution of global players, whereas Topic GPT captures more detailed thematic relationships.

The differences highlight the complementary nature of LDA and Topic GPT in thematic analysis.

Word Clouds for Clusters: Plotting word clouds for the two clusters identified by both LDA and Topic GPT provides a visual representation of the predominant themes within each cluster. This allows for an intuitive comparison of the thematic content and differences between clusters, highlighting the distinct focus areas captured by each model. Word clouds help in quickly identifying key terms and topics, facilitating a deeper understanding of the thematic structure within each cluster. This visual approach complements quantitative

analysis, offering a comprehensive view of the data and aiding in the interpretation of complex textual information.

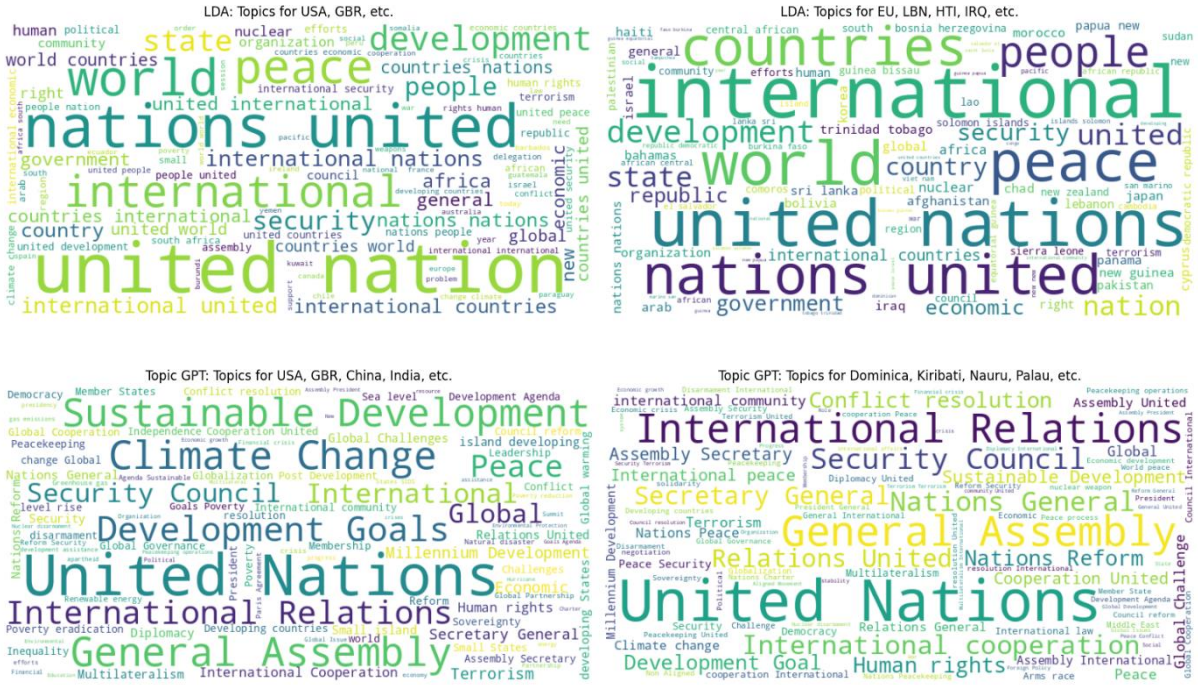

Figure 7: Word Clouds for two Clusters from LDA and TopicGPT

LDA Clustering Interpretation

Cluster 1 (left word cloud):

- Countries Included: USA, GBR, IND, RUS, CHN, UKR, etc.
- Dominant Themes: This cluster prominently features terms such as "united," "nations," "international," "peace," "development," and "world." The recurring focus on "peace" and "development" indicates a strong emphasis on global stability and progress. The presence of terms like "security," "economic," and "government" suggests discussions around governance and economic issues.

Cluster 2 (right word cloud):

- Countries Included: EU, LBN, HTI, IRQ, SDN, etc.

- Dominant Themes: This cluster also highlights terms like "united," "nations," "international," and "peace," but with more emphasis on "countries," "development," "people," and "world." The frequent mention of "countries" and "people" suggests a focus on social and regional issues. The themes of "development" and "economic" indicate ongoing discussions about regional development and economic growth.

Topic GPT Clustering Interpretation

Cluster 1 (left word cloud):

- Countries Included: USA, GBR, China, India, EU, etc.
- Dominant Themes: This cluster shows a strong emphasis on "United Nations," "General Assembly," "International Relations," "Security Council," "Climate Change," and "Sustainable Development." The inclusion of terms like "Millennium Development Goals" and "Global Governance" reflects a focus on comprehensive global initiatives and governance. The frequent mention of "peace" and "international cooperation" indicates ongoing dialogues about global peace and collaborative efforts.

Cluster 2 (right word cloud):

- Countries Included: Dominica, Kiribati, Nauru, Palau, etc.

- Dominant Themes: This cluster highlights "United Nations," "International Relations," "Security Council," and "General Assembly," similar to Cluster 1. However, there is a stronger focus on "development goals," "human rights," "peacekeeping," and "conflict resolution." The recurring mention of "international cooperation" and "terrorism" points to discussions around security and collaborative efforts to address global challenges.

Comparative Analysis: Both LDA and Topic GPT highlight the significance of the United Nations, international relations, and peace across different clusters. However, there are notable differences in the thematic focus of the clusters identified by each model.

LDA clusters broadly capture themes around global cooperation, peace, development, and governance, with a mix of high-level international topics and regional developmental issues. In contrast, Topic GPT provides a more detailed and nuanced understanding, capturing specific initiatives such as "Climate Change," "Millennium Development Goals," and "Human Rights," reflecting deeper thematic relationships and more targeted discussions.

These visualizations illustrate the strengths of each model: LDA provides a broad overview of recurring themes, while Topic GPT delves into detailed and specific thematic structures. This combined approach enhances our understanding of the thematic diversity and priorities within UNGA speeches, offering valuable insights for policymakers and researchers.

Conclusion

The analysis of United Nations General Assembly (UNGA) speeches from 1971 to 2018 using Latent Dirichlet Allocation (LDA) and Topic GPT has provided significant insights into the thematic structures and evolution of international discourse over nearly five decades. By leveraging both LDA and Topic GPT, we were able to capture a comprehensive and nuanced understanding of the topics discussed in these speeches, highlighting the strengths and complementary nature of these two advanced topic modeling techniques.

Enhanced Thematic Coverage: The application of LDA offered a broad overview of the major themes discussed in the UNGA speeches. LDA's ability to capture general distributions of topics across a large corpus was instrumental in identifying overarching trends in the speeches. Key themes such as peace, development, security, and international cooperation emerged as central topics, reflecting the UN's enduring focus on global stability and progress.

In contrast, Topic GPT, with its advanced contextual embeddings, provided deeper and more nuanced insights into specific entities and relationships within the text. This model's ability to capture semantic relationships between words allowed for the identification of detailed themes such as climate change, sustainable development, human rights, and terrorism. The inclusion of specific initiatives like the Millennium Development Goals further highlighted the UN's targeted efforts to address global challenges.

Improved Analytical Accuracy: The integration of both LDA and Topic GPT significantly enhanced the analytical accuracy of our thematic analysis. While LDA provided a statistically robust framework for understanding general themes, Topic GPT's contextual richness ensured that the clusters formed were not only thematically similar but also contextually relevant. This dual-model approach allowed for a more accurate and coherent clustering of topics, reflecting the true thematic structures within the speeches.

Comprehensive Cluster Formation: By clustering the topics derived from both LDA and Topic GPT, we achieved a balanced and multi-dimensional understanding of the thematic landscape. LDA's broad themes offered macro-level insights, while Topic GPT's detailed themes provided micro-level granularity. This combination ensured that the clusters formed were comprehensive and interpretable, accurately representing the thematic diversity of the speeches. The clustering results demonstrated clear distinctions between the thematic focuses of different groups of countries, highlighting both general trends and specific issues pertinent to various regions and nations.

Visual Insights and Comparative Analysis: The use of word clouds to visualize the top 100 topics from both models provided an intuitive and immediate understanding of the thematic content. These visualizations highlighted the central themes and key terms within each cluster, allowing for a quick comparison of the thematic differences and similarities between

the models. LDA's word clouds showed a broad focus on international cooperation and development, while Topic GPT's word clouds revealed more detailed and nuanced themes, capturing specific initiatives and deeper relationships.

Strategic Implications for Policymakers and Researchers: The insights gained from this project have significant implications for both policymakers and researchers. For policymakers, understanding the thematic evolution and priorities within UNGA speeches can inform the development of more targeted and effective policies. The detailed themes captured by Topic GPT, such as climate change and sustainable development, highlight areas that require ongoing international cooperation and focus. Researchers can benefit from the comprehensive thematic analysis provided by both models, identifying areas for further investigation and understanding the historical context of international relations.

Final Thoughts: While LDA and Topic GPT both serve as powerful tools for thematic analysis, each has distinct advantages. LDA is highly efficient and excels in providing a broad overview of major themes, making it suitable for quickly identifying overarching trends across large text corpora. Its probabilistic framework ensures robust and generalizable insights, particularly useful for high-level thematic exploration.

Topic GPT, on the other hand, offers significant advantages in capturing detailed and nuanced themes due to its use of advanced contextual embeddings. This model excels in identifying specific entities and relationships, providing richer and more granular insights into the text. Its ability to capture semantic relationships between words makes it particularly effective for in-depth thematic analysis, revealing complex themes that might be overlooked by LDA.

In conclusion, the complementary strengths of LDA and Topic GPT make them valuable tools for a comprehensive thematic analysis. While LDA provides a solid foundation for understanding broad thematic trends, Topic GPT enhances this analysis by delving deeper into the nuances and specifics of the text. This integrated approach offers a more holistic and insightful understanding of the thematic structures within UNGA speeches, reflecting the evolving priorities and challenges of the international community over nearly five decades.

References

1. Blei, David & Ng, Andrew & Jordan, Michael. (2001). Latent Dirichlet Allocation. The Journal of Machine Learning Research. 3. 601-608. (Link)
2. Rahul Kumar Gupta, Ritu Agarwalla, Bukya Hemanth Naik, Joythish Reddy Evuri, Apil Thapa, Thoudam Doren Singh, Prediction of research trends using LDA based topic modeling. (Link)
3. A. Goyal and I. Kashyap, Latent Dirichlet Allocation - An approach for topic discovery. (Link)
4. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (Link)
5. Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, Mohit Iyyer. TopicGPT: A Prompt-Based Topic Modeling Framework. (Link)