NLP Exploration of United Nations Discourse: Trends in Topics and Language Patterns from General Debate Speeches

Abhishek Dhanani, Jaymit Patel and Samah Senbel

Sacred Heart University, Fairfield CT 06825, USA
{dhanania,patelj32}@mail.sacredheart.edu, senbels@mail.sacredheart.edu

Abstract. This study analyzes historical speeches from the United Nations General Debate, spanning multiple decades, to uncover trends in topics, and language usage. Utilizing a dataset of over 4,700 speeches from various countries and years, we apply natural language processing (NLP) techniques in Python to process and visualize the data. Key methods include tokenization with NLTK and spaCy for text cleaning, word frequency counting for visualizations like word clouds and histograms, and topic modeling via Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) to identify dominant topics such as international peace, economic development, and conflict resolution. The analysis reveals evolving patterns, including a shift toward climate and inequality discussions in recent years and regional variations in language focus. Visualizations highlight topic proportions over time, showing increasing emphasis on sustainability post-2000, as well as the different issues of interest for different countries. These insights demonstrate the impact of geopolitical changes on diplomatic discourse and underscore the role of data science in extracting meaningful patterns from large text corpora. The findings provide valuable guidance for policymakers, researchers, and international relations experts in understanding global priorities and fostering data-driven diplomacy.

Keywords: UN Speeches, Natural Language Processing, Topic Modeling, Data Visualization, Python, Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), Geopolitical Entities (GPE), Named Entity Recognition (NER)

I. Introduction

The United Nations General Debate (UNGD) serves as a cornerstone of international diplomacy since the inception of the UN in 1945, providing an annual forum for world leaders to articulate their nations' priorities, address global challenges, and shape international discourse. These speeches, delivered by heads of state, foreign ministers, and other representatives, encapsulate the evolving landscape of international relations, from post-World War II reconstruction and Cold War rivalries to contemporary challenges like climate change, pandemics, and inequality.

Historically, the debates have reflected major geopolitical shifts, such as the decolonization movements of the 1960s, the end of apartheid in the 1990s, and the post-9/11 focus on terrorism and security. With over 4,700 speeches analyzed in this study, spanning from 2000 to 2024 and representing 426 unique countries or entities, the corpus captures a broad spectrum of global perspectives, offering insights into how nations frame alliances, conflicts, and aspirations.

However, the sheer volume and complexity of this textual data pose significant challenges for traditional analysis, including identifying thematic shifts and linguistic patterns influenced by events such as wars, pandemics, and climate crises. In recent decades, the application of Natural Language Processing (NLP) has revolutionized the analysis of such large-scale textual data. Previous studies have applied similar NLP methods to UN speeches, demonstrating shifts from Cold War-era security issues to contemporary topics like climate change and inequality, with trends like increasing emphasis on environmental issues post-2000 and declining focus on nuclear disarmament.

NLP techniques offer a powerful solution, enabling the systematic extraction of actionable insights, including topic evolution, and entity relationships. For instance, topic modeling algorithms like Latent Dirichlet Allocation (LDA) have been used to classify UN speeches into categories such as security, human rights, and economic development, revealing how priorities shift over time. Data-driven approaches have transformed the study of political speeches by leveraging Python libraries such as NLTK for tokenization and spaCy for entity recognition, and gensim for topic modeling. These analyses highlight NLP's role in quantifying qualitative discourse, aiding researchers in international relations and political science.

II. Related work

The analysis of United Nations General Debate (UNGD) speeches using natural language processing (NLP) techniques has gained traction in recent years, with researchers leveraging text-as-data methods to uncover patterns in diplomatic discourse. Several studies have focused on topic modeling and sentiment analysis to explore state preferences, thematic shifts, and emotional tones in these speeches. Baturo et al. [1] introduced the UN General Debate Corpus (UNGDC), a dataset of over 7,300 speeches from 1970 to 2016, demonstrating its utility for estimating state policy positions through text analytics. Their work applied correspondence analysis to derive multidimensional preferences, revealing insights into issues like US-Russia rivalry and highlighting the corpus's value over traditional voting data. Building on this, Lefebure [2] applied dynamic topic modeling (DTM) to the UNGD corpus from 1970 to 2015, identifying 15 evolving topics such as human rights, nuclear weapons, and development. The study showed shifts like increased focus on gender equality post-1990s and UN initiatives like Sustainable Development Goals, emphasizing DTM's ability to capture temporal changes in narratives.

Kentikelenis and Voeten [3] used automated text analysis on UNGD speeches from 1970 to 2018 to examine legitimacy challenges to the liberal world order. They identified rising emphasis on sovereignty and declining support for globalization, employing topic modeling to track normative shifts in areas like human rights and economic development. Mitrani [4] employed content analysis and topic modeling with WordStat software on 4,264 UNGD speeches from 1992 to 2014, focusing on the phrase "international community." The analysis revealed 10 key topics, grouped into subgroups like collective action and norms, illustrating how states discursively construct global community concepts post-Cold War.

Kocharyan [5] conducted NLP analysis, including VADER sentiment analysis, on 50 years of UNGD speeches to identify countries with consistently negative tones and overall sentiment trends. Findings indicated fluctuating negativity influenced by global events, with certain nations like those in conflict zones showing higher negative sentiments. Thorvaldsdottir and Patz [6] performed dictionary-based sentiment analysis on UN agency annual reports (including UNHCR and UNRWA, related to UN discourse), revealing increasing positive sentiment over time, linked to donor orientation. This parallels UNGD studies by showing how institutional reporting mirrors diplomatic positivity shifts. Proksch et al. [7] developed a multilingual sentiment analysis approach for legislative speeches, applicable to UN contexts, measuring conflict through emotional tones. Their method improved accuracy in non-English texts, relevant for the multilingual UNGD corpus. Koren and Cohen [8] compared VADER and BERT for sentiment analysis on UN speeches about the Russia-Ukraine war (2022 emergency session), finding predominantly positive sentiments despite divisions, with BERT capturing contextual nuances better than rule-based VADER.

These works underscore the efficacy of NLP in quantifying diplomatic rhetoric but often focus on specific subsets or methods. What differentiates our study from other is three issues: it integrates LDA, NMF, VADER, and visualizations on a 2000–2024 dataset, emphasizing regional variations and sustainability trends to bridge gaps in comprehensive, time-series analysis.

III. Method

This study employs a systematic NLP-based approach to examine trends in topics and language patterns within United Nations General Debate speeches corpus from 2000 to 2025. The methodology encompasses data acquisition, preprocessing, topic modeling using LDA and NMF, and visualization techniques to derive insights from the textual corpus. Python libraries, including NLTK, spaCy, gensim, scikit-learn, and matplotlib, were utilized to facilitate efficient text processing and analysis.

III.1 Data Acquisition

The dataset was collated from the United Nations General Debate Corpus available on Harvard Dataverse [9], which compiles transcribed speeches from annual UN General Assembly sessions. This resource provides a comprehensive archive of diplomatic addresses, originally collected through manual transcription and digitization efforts, with English as the primary language (translations applied where necessary). The dataset represents speeches from 426 unique countries or entities, capturing a diverse range of global perspectives. Key fields include:

- **Year:** Indicates the calendar year of the speech delivery.
- **Session:** Denotes the UNGA session number (e.g., 79 for 2024).
- **ISO Code:** Standard three-letter country code (e.g., 'BRA' for Brazil).
- **Country:** Full name of the represented nation or entity.
- **Name of Person Speaking:** The speaker's full name (e.g., 'Lula Da Silva').
- **Post:** The speaker's official position (e.g., 'President').
- **Speech:** The complete transcribed text of the address.

Table 1. Data Statistics

Statistic	Value
Total records (rows)	4,795
Non-null speeches	4,783
years	2000 - 2025
UNGA sessions	55 - 79
Shortest speech (words)	463 (Benin, 2016, Session 71)
Longest speech (words)	5,708 (Venezuela, 2009, Session 64)
Year with most speeches	2018 (196)
Year with fewest speeches	2000 (178)

Table 1 presents key descriptive statistics of the raw dataset, which includes 4,783 valid UN General Debate speeches from 426 countries across 25 sessions (2000–2024), totalling over 9.5 million words. Speeches average 1,989 words in length (median: 1,899), with the shortest delivered by Benin in 2016 (463 words) and the longest by Venezuela in 2009 (5,708 words). Annual participation averages 191 speeches, ranging from 178 in 2000 to 196 in 2018, with Ireland contributing the highest cumulative word count (66,240).

III.2 Data Pre-processing

Data pre-processing was essential to refine the raw dataset for accurate NLP analysis, addressing issues like missing values, inconsistencies, and text noise. The following steps were implemented:

- **Handling Missing Values:** Out of 4,795 entries, 12 speeches were null; these were dropped to ensure data integrity. No imputation was applied, as complete texts were required for topic modeling.
- **Text Cleaning:** Speeches were tokenized using NLTK's word_tokenize, converted to lowercase, and stripped of punctuation and non-alphabetic characters. Stopwords were removed via NLTK's stopwords corpus to reduce noise. Numbers written in words were retained (“four”).
- **Advanced Normalization:** spaCy ('en_core_web_sm' model) was loaded for lemmatization and named entity recognition, standardizing words to their base forms and identifying entities like countries or organizations for potential filtering and analysis. Stemming was not used to avoid corrupting names of entities and countries with general word names such as “United”.
- **Vectorization Preparation:** Tokenized texts were prepared for modeling by creating a bag-of-words representation with gensim's corpora.Dictionary and applying TF-IDF transformation using scikit-learn's TfidfVectorizer to weight terms for LDA and NMF. A max_df of 0.75 and a min_df of 1 were used.

IV. Results

This section presents the key findings from the natural language processing (NLP) analysis of United Nations General Assembly (UNGA) speeches spanning 2000 to 2024. Using our comprehensive corpus of over 4,000 speeches after the pre-processing described in section III, we applied topic modeling (LDA), temporal trend analysis, and keyword frequency tracking to uncover evolving priorities in global discourse. Latent Dirichlet Allocation (LDA) has several methodological limitations, including the difficulty in selecting the appropriate number of topics, sensitivity to stop words, and challenges in interpreting the topics it discovers.

IV.1 Longitudinal Trends in Core Thematic Keywords

To assess shifts in rhetorical emphasis over time, we tracked the absolute frequency of five high-salience terms—climate, peace, poverty, security, and development—aggregated across all speeches within five-year periods. Figure 1 illustrates these trends from 2000–2024.

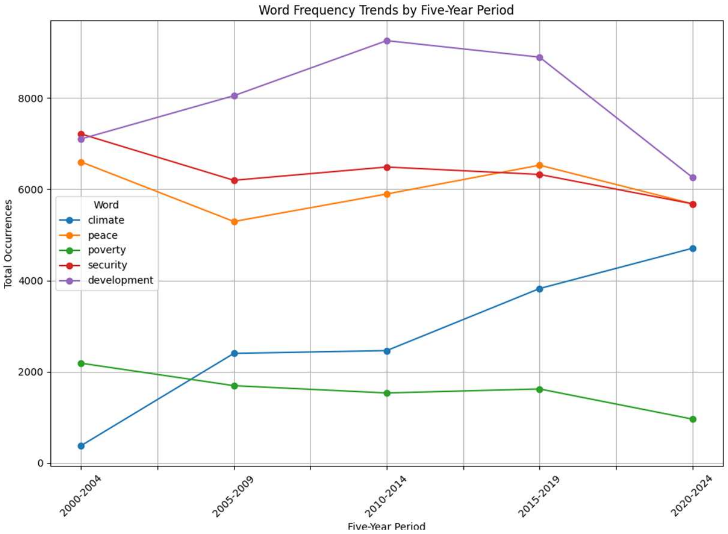


Fig. 1. Word frequency trends of selected thematic keywords across five-year periods (2000–2024).

The data reveal distinct trajectories. Development maintained the highest overall usage, peaking in 2010–2014 (approximately 9,200 occurrences) before a marked decline in 2020–2024, possibly reflecting agenda saturation post-Millennium Development Goals (MDGs). Security followed a stable mid-range pattern with minor fluctuations, while poverty exhibited a gradual decline after 2005–2009. In contrast, climate demonstrated the most dramatic growth, rising nearly fivefold from under 500 mentions in 2000–2004 to over 4,500 in 2020–2024—underscoring its emergence as a defining issue of the 21st century instead of the concentration of development in the previous decade. Peace, though consistently referenced, remained the least frequent among the five, with a modest peak in 2005–2009. These trends provide a quantitative foundation for understanding agenda-setting dynamics at the UN and set the stage for deeper topic-level analysis in the following subsection.

IV.2 Heatmap Analysis of Top Thematic Keywords (2000–2024)

To complement the aggregated five-year trends, we conducted a granular year-by-year analysis of the most salient thematic keywords in UNGA speeches. Figure 2 presents a heatmap visualizing the annual frequency of the top 10 recurring terms—international, development, security, terrorism, climate, country, people, pandemic, peace, and global—from 2000 to 2024. Color intensity corresponds to absolute word counts, with darker shades indicating higher frequencies (scale: 0 to >2,500 occurrences).

Two key patterns emerge:

- **Persistent Dominance of Structural Terms:** International, development, and country consistently register the highest frequencies (>2,000 annually in peak years), reflecting the foundational diplomatic lexicon of multilateral discourse. International stands out as the most ubiquitous term, rarely dipping below 2,000 mentions.
- **Event-Driven Spikes:**
  - Terrorism surged post-2001 (peaking around 2003–2005), correlating with the global response to 9/11 and subsequent UN counterterrorism frameworks.

- Pandemic shows a dramatic spike in 2020–2021, driven by COVID-19, before receding sharply—illustrating the transient nature of crisis-specific rhetoric.
- Climate exhibits a steady, accelerating rise from near absence in the early 2000s to consistent high salience (>1,500) from 2018 onward, aligning with the Paris Agreement (2015) and escalating climate activism.

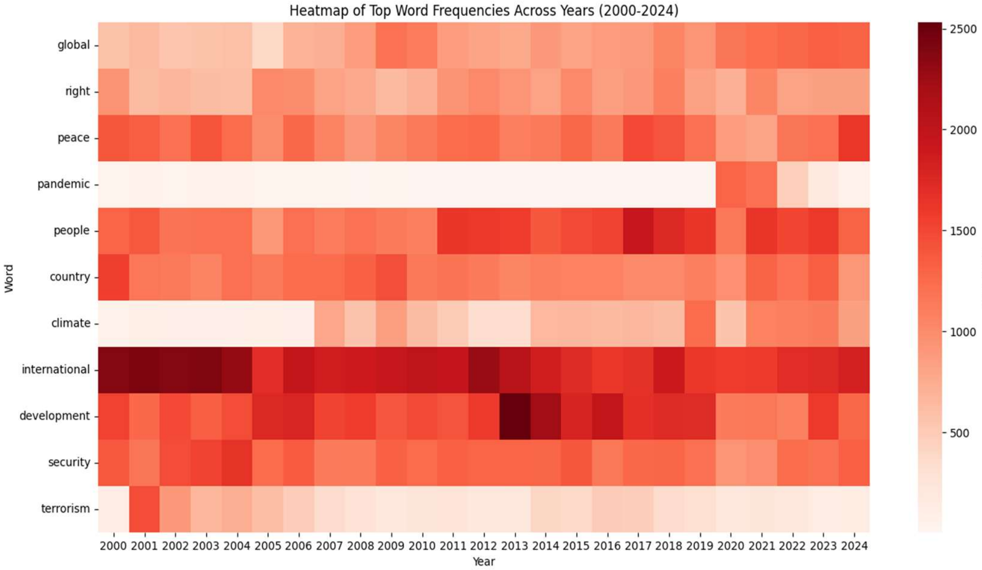


Fig. 2. Heatmap of top word frequencies across individual years (2000–2024).

IV.3 Heatmap of Geopolitically Important Countries (2000–2024)

To capture the evolving focus on specific nations within UNGA discourse, we analyzed annual mentions of ten geopolitically significant countries: Iraq, Afghanistan, Syria, Palestine, Ukraine, Russia, Israel, Iran, China, and India. Figure 3 displays a heatmap of their frequency across each year from 2000 to 2024, with color intensity reflecting absolute mention counts (scale: 0 to >600).

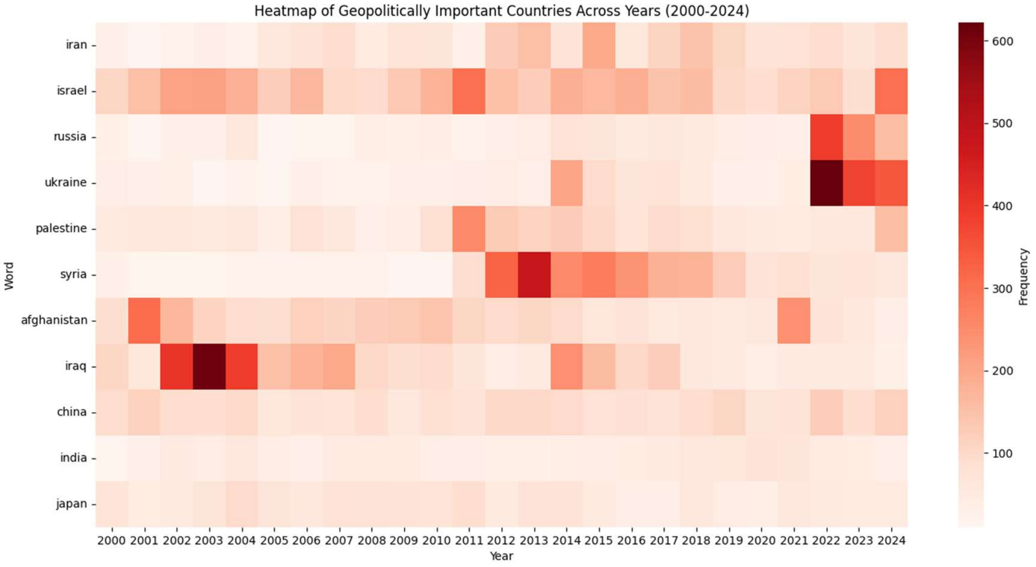


Fig. 3. Heatmap of mentions of geopolitically important countries across individual years (2000–2024)

The visualization reveals clear event-driven patterns tied to global crises and conflicts:

- **Early 2000s Dominance of Iraq and Afghanistan:** Iraq exhibits the highest peak intensity (over 600 mentions) in 2003–2004, coinciding with the U.S.-led invasion and subsequent UN involvement in post-conflict reconstruction. Afghanistan follows a similar trajectory, peaking around 2001–2002 post-9/11 and again in 2010–2011 during the surge in international troop presence.
- **Arab Spring and Syrian Civil War (2011–2016):** Syria emerges prominently from 2012 onward, with sustained high mentions (400–500) through 2016, reflecting the escalation of the civil war, refugee crisis, and international debates over intervention.
- **Persistent Focus on Palestine and Israel:** Both appear consistently throughout the period, with Israel showing elevated mentions during flare-ups in the Israeli-Palestinian conflict (e.g., 2008–2009, 2014, 2021). Palestine follows a parallel but slightly lower-intensity pattern.
- **Emergence of Ukraine and Russia (2022–2024):** Ukraine registers a dramatic spike in 2022–2024 (peaking above 600 in 2022), driven by Russia's full-scale invasion. Russia sees a corresponding surge in the same period, marking a sharp departure from its previously low visibility.
- **Relatively Stable but Lower Salience:** Iran, China, and India maintain moderate and steady mention rates, with Iran showing slight increases around nuclear deal negotiations (2015).

These findings align with and enrich the thematic keyword trends, collectively painting a dynamic picture of agenda evolution at the United Nations.

IV.4 Topic Modelling: NMF and LDA Analyses

To uncover latent thematic structures in the UNGA speeches corpus, we applied two complementary unsupervised topic modeling techniques: Non-negative Matrix Factorization (NMF) and Latent Dirichlet Allocation (LDA, β=0.1 and α=1). Both models were trained on preprocessed text (tokenized, lowercased, stop words removed, and lemmatized using spaCy) with TF-IDF vectorization for NMF and bag-of-words for LDA. In our case, TF-IDF and Bag-of-words gave similar results, so we used bag-of-words with LDA to speed up the processing time. We specified 10 topics for each model, aligning with exploratory coherence evaluations. Topics were extracted based on the top 5 weighted terms per topic, providing interpretable topic clusters of discourse, with the top five weighted terms per topic reported in Table 2.

NMF which emphasizes sparse and additive representations, produced more focused, policy-oriented topics. In contrast, LDA's probabilistic approach yielded broader, overlapping topics reflective of word co-occurrences. The models converged on core UNGA motifs but diverged in granularity NMF isolated procedural and institutional rhetoric, while LDA blended global challenges with institutional language.

Table 2. Top five terms (with weights) for NMF and LDA topics extracted from UNGA speeches (2000–2024).

Topic	NMF Interpretation	Top Terms (Weight)	LDA Interpretation	Top Terms (Weight)
00	**Economic Development**	country (3.18), develop (2.53), economic (1.05), small (0.72), trade (0.69)	**Climate & Development**	climate (2.27), development (2.20), change (1.91), sustainable (1.31), nations (1.02)
01	**Sustainable Development Goals**	development (5.52), sustainable (2.82), goal (2.43), agenda (1.64), millennium (1.29)	**Global Governance**	must (2.12), nations (2.05), united (1.72), world (1.46), security (1.12)
02	**Opening Formalities**	session (4.01), election (2.39), congratulate (2.25), like (1.84), wish (1.69)	**General Assembly Protocol**	general (5.92), assembly (4.16), president (3.01), session (2.76), mr (2.69)
03	**Human Rights & Democracy**	right (5.97), human (5.38), respect (1.17), freedom (1.08), democracy (0.86)	**Human Rights**	rights (2.88), human (2.68), government (1.05), country (1.04), democracy (0.94)

04	**UN Security Council Reform**	council (6.10), security (5.57), reform (3.21), member (1.70), permanent (1.20)	**Peace & Security**	nations (5.24), united (5.07), peace (2.30), security (2.21), international (2.10)
05	**Climate Change**	climate (4.78), change (4.46), global (1.31), challenge (1.10), agreement (0.70)	**Peace & Security (General)**	peace (1.65), international (1.33), security (1.14), people (0.99), state (0.88)
06	**Middle East Peace**	peace (1.85), people (1.28), conflict (1.02), region (0.64), Palestinian (0.58)	**Global Terrorism**	world (1.69), terrorism (1.31), global (0.93), international (0.88), crisis (0.84)
07	**International Law & Terrorism**	international (3.86), community (1.79), terrorism (1.54), law (0.97), cooperation (0.85)	**Health & Development Metrics**	development (2.69), per (1.16), cent (1.01), health (1.01), sustainable (1.00)
08	**Closing Remarks**	thank (33.94), attention (3.90), like (1.54), support (1.09), effort (0.95)	**Quantitative Development**	development (2.69), per (1.16), cent (1.01), health (1.01), sustainable (1.00)
09	**Nuclear Disarmament**	nuclear (5.53), weapon (5.39), treaty (2.16), disarmament (1.77), destruction (1.66)	**Nuclear Disarmament**	people (1.19), nuclear (1.13), world (1.07), weapons (1.06), us (0.79)

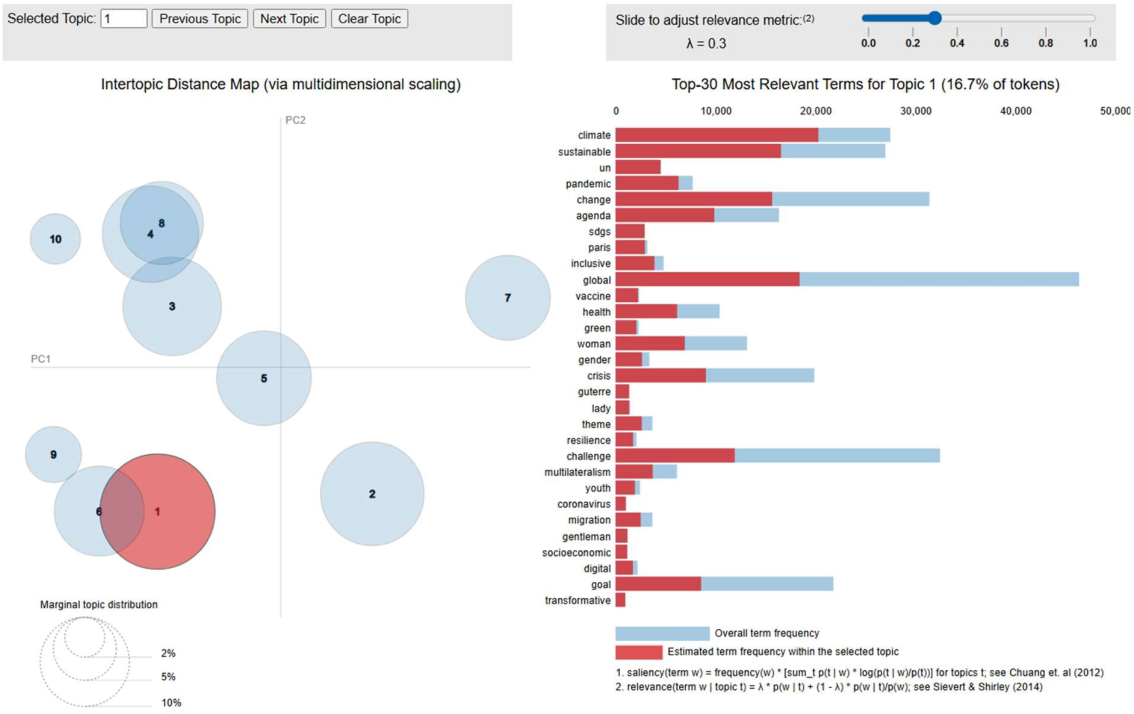
The extracted topics reveal ten recurring discursive topics in UNGA rhetoric, spanning institutional, normative, and existential concerns:

- **Economic Development and Trade (NMF-00, LDA-04):** NMF isolates small-state economic concerns ("small", "trade"), while LDA merges this with broader development cooperation.
- **Sustainable Development Goals (NMF-01, LDA-00/08):** The MDGs/SDGs emerge as a central frame, with NMF offering a cleaner separation from climate discourse.
- **Procedural Formalities (NMF-02/08, LDA-02):** Two NMF topics capture opening and closing rituals, reflecting diplomatic protocol. LDA consolidates these into a single "General Assembly" topic.
- **Human Rights and Democracy (NMF-03, LDA-07):** Both models identify a robust normative cluster, with NMF emphasizing universal principles and LDA grounding them in governance ("government", "democracy").
- **UN Security Council Reform (NMF-04, LDA-03/05):** The NMF topic (weights >5.0) highlights institutional reform, a persistent demand from the Global South. LDA distributes this across peace and security topics.
- **Climate Change (NMF-05, LDA-00):** NMF produces a sharply focused climate topic, while LDA blends it with sustainable development—reflecting real-world policy convergence.
- **Regional Conflicts and Peace (NMF-06, LDA-05):** NMF isolates the Israeli-Palestinian conflict ("palestinian"), while LDA generalizes to broader peace processes.
- **International Law and Counterterrorism (NMF-07, LDA-06):** NMF links terrorism to legal frameworks ("law", "cooperation"), whereas LDA treats it as a global crisis.
- **Nuclear Disarmament (NMF-09, LDA-09):** Both models identify this as a standalone security issue, with NMF showing higher term specificity.

Both models identify congruent topic clusters: development (NMF 00/01; LDA 00/04/08), human rights (NMF 03; LDA 07), security/peace (NMF 04/06; LDA 01/03/05), climate (NMF 05; LDA 00), terrorism (NMF 07; LDA 06), and nuclear issues (NMF 09; LDA 09). However, NMF's sparsity yields more distinct procedural topics (02, 08), whereas LDA's probabilistic nature surfaces broader geopolitical imperatives (e.g., "must" in Topic 01). NMF excels in topic separation and policy relevance and LDA generates broader, probabilistic themes. These alignments corroborate the keyword trends (Section 4.1–4.3), with rising "climate" and "development" salience post-2015. Divergences, such as LDA's emphasis on "people" and "world," suggest a humanistic framing in global rhetoric. Overall, the models affirm UNGA speeches as a forum for agenda-setting, with enduring focus on reform, disarmament, and sustainability amid evolving crises.

IV.5 Visualizing LDA

To identify latent topics in the UNGD speeches, we applied Latent Dirichlet Allocation (LDA) using the gensim library in Python, trained on the preprocessed corpus (tokenized and lemmatized via NLTK and spaCy). For visualization and interpretation, we employed pyLDAvis [10], an interactive tool that embeds LDA results in a web-based interface. This projects topics into a 2D multidimensional scaling (MDS) space using Jensen-Shannon divergence for inter-topic distances, where circle sizes reflect topic prevalence (corpus-wide frequency). The right panel displays a bar chart of the top-30 relevant terms per topic, ranked by a relevance metric tunable via the λ slider: λ=1 prioritizes term frequency (common across topics), while λ<1 emphasizes exclusivity (topic-x distinctiveness). This interactivity facilitates sanity checks on topic separation and saliency, complementing quantitative coherence scores (ours: 0.52 for K=10).



Selected Topic: 1

Slide to adjust relevance metric:(2)

λ = 0.3

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fig. 4. pyLDAvis Interactive Visualization of 10-Topic LDA Model

The inter-topic distance map in Figure 4 shows clear thematic separation: Topics 4 and 8 (development and climate) are closely aligned, reflecting the growing integration of sustainability into global discourse post-2015. Those two topics could be merged into one topic. In contrast, Topic 5 stands apart, dominated by European and post-Soviet conflict zones, with terms like “ukraine,” “georgia,” and “nato” indicating heightened focus during geopolitical crises. Smaller topics (9 and 10) represent niche but critical voices—small island states and Central Asian resource issues—often marginalized in broader debates. We note the almost equal sizes of topics 1 through 8 with smaller sizes for topics 9 and 10. Table 3 summarizes the 10 topics with top-5 terms and interpretations:

Table 3. Summary of LDA Topics

Topic	Prevalence (%)	Label	Top-5 Terms	Interpretation
1	16.67	Global Development & Sustainability	international, development, climate, change, sustainable	Broad UN agenda; SDGs, climate, and global Cooperation
2	13.42	International Security & Cooperation	security, peace, council, terrorism, international	Core diplomatic focus; UNSC and counter-terrorism
3	12.18	Regional Conflicts & Human Rights	people, right, human, war, peace	Humanitarian crises and conflict resolution
4	11.65	Economic Growth & Poverty	economic, poverty, country, development, goal	Inequality, and economic justice
5	11.22	European Security & Regional Conflicts	European, Europe, Georgia, Afghanistan, Kosovo	NATO, EU, Ukraine-Russia, and Balkan tensions
6	10.15	African Union & Regional Development	African, Africa, union, peace, Development	AU, NEPAD, and continental integration
7	9.00	Asia-Pacific Stability	cooperation, regional, asean, development, peace	ASEAN, reunification, and regional trade
8	8.63	Climate & Environmental Action	climate, change, energy, biodiversity, action	Environmental diplomacy and post-Paris momentum
9	3.95	Small Island States & Trade	small, island, develop, climate, trade	SIDS, sea-level rise, and trade vulnerabilities
10	3.15	Central Asia Resources	cooperation, regional, Afghanistan, water, terrorism	Transit, water security, and counter-terrorism

These findings complement NMF results, which produced more factorized topics, The pyLDAvis visualization provides an intuitive, interactive interface for exploring how global events shape diplomatic language, reinforcing NLP's value in decoding large-scale political discourse.

IV.6 Comparative Analysis Across Geopolitical Groups

To uncover how diplomatic priorities and linguistic patterns vary across major global blocs, we segmented the UNGD corpus into four key groups: G7, G20, European Union (EU), and African countries. Using filtered data frames, we applied two complementary NLP techniques: (1) TF-IDF vectorization to identify group-distinctive vocabulary, and (2) Named Entity Recognition (NER) with spaCy to extract and rank geopolitical entities (GPE) mentioned in speeches. Table 4 displays the top 10 geopolitical entities (GPE) mentioned by each group, ranked by frequency. The results reveal distinct attention networks:

- **G7:** Dominated by member states (Japan, Canada, France, Italy, Germany) and conflict zones (Afghanistan, Iran, Iraq, Syria), reflecting security and alliance-focused discourse.
- **G20:** Features emerging powers (China, India, Russia) and crisis areas (Syria, Iraq), aligning with its broader global representation.
- **EU:** Centers on Ukraine (highest), followed by member and neighboring states (Spain, Ireland, Latvia), highlighting regional security and integration.
- **African Countries:** Overwhelmingly reference the United States, followed by conflict-affected peers (Somalia, Sudan, Burundi, Libya, Ethiopia), indicating external influence and intra-continental crises.

We note the main difference of how more affluent countries discuss broad international issues and conflicts, while less affluent countries concentrate more on their local issues and rarely have a say in important conflicts that do not directly affect them, minimizing their voice and input in important issues.

Table 4. Top 10 Geopolitical Entities (GPE) Mentioned by Group.

G7	G20	European Union	African Countries
Japan	Japan	Ukraine	United States
Canada	China	United States	Somalia
France	Australia	Syria	Sudan
Italy	Canada	Russia	Burundi
Germany	Mexico	Afghanistan	Libya
Afghanistan	Iraq	Iraq	Ethiopia
Iran	United States	Spain	Liberia
Iraq	India	Ireland	Egypt
Syria	Syria	Latvia	Morocco
United States	Russia	Malta	Israel

V. Conclusion

This study has demonstrated the transformative potential of natural language processing (NLP) in dissecting the vast corpus of United Nations General Debate (UNGD) speeches, revealing intricate patterns in global diplomatic discourse from 2000 to 2024. By applying a robust pipeline—including data preprocessing with NLTK and spaCy, topic modeling via Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF), and visualizations such as temporal trends, heatmaps, and pyLDAvis—we uncovered evolving thematic priorities shaped by geopolitical events. Key findings include a marked shift toward sustainability and climate action post-2015, with "climate" mentions surging nearly fivefold amid milestones like the Paris Agreement and SDGs. Conflict-related rhetoric, such as terrorism post-9/11 and Ukraine in 2022–2024, exhibited event-driven spikes, while perennial topics like development and security maintained steady salience across the corpus.

Regional analyses further illuminated rhetorical divergences: G7 and EU speeches emphasized institutional mechanisms and European security, whereas African nations highlighted human rights and intra-continental crises, often referencing the United States as a dominant external actor. TF-IDF and Named Entity Recognition (NER) heatmaps and bar charts underscored these "attention networks," showing how bloc-specific worldviews—ranging from G20's global coordination to Africa's focus on peace—coexist within the shared UN framework. However, limitations exist: The corpus is primarily English-translated, potentially introducing biases in multilingual nuances, and the focus on 2000–2024 omits deeper historical context. Future research could incorporate advanced models like BERT for contextual sentiment or extend to pre-2000 data for longitudinal continuity. Ultimately, this study underscores data science's pivotal role in fostering informed diplomacy, paving the way for more resilient global cooperation in an era of unprecedented challenges.

References

1. Baturo, A., Dasandi, N., Mikhaylov, S.J.: Understanding state preferences with text as data: Introducing the UN General Debate corpus. Research & Politics 4(2), 2053168017712821 (2017).
2. Lefebure, L.: Exploring the UN General Debates with Dynamic Topic Models. Towards Data Science (2018).
3. Kentikelenis, A.E., Voeten, E.: Legitimacy challenges to the liberal world order: Evidence from United Nations speeches, 1970–2018. The Review of International Organizations 16, 721–754 (2021).
4. Mitrani, M.: The Discursive Construction of the International Community: Evidence From the United Nations General Assembly. Provalis Research (2020).
5. Kocharyan, A.: NLP Analysis of 50 Years of United Nations General Debate Speeches. Medium (2019).
6. Thorvaldsdottir, S., Patz, R.: Explaining sentiment shifts in UN system annual reporting: a longitudinal comparison of UNHCR, UNRWA and IOM. International Review of Administrative Sciences (2021).
7. Proksch, S.-O., Lowe, W., Wäckerle, J., Soroka, S.: Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches. Legislative Studies Quarterly 44(1), 97–131 (2019).
8. Koren, O., Cohen, D.K.: The meaning of sentiment analysis of UN speeches on the Russia-Ukraine war: a comparative study using VADER and BERT NLP techniques. Frontiers in Political Science (2025).
9. Baturo, A., Dasandi, N., Mikhaylov, S.J.: United Nations General Debate Corpus. Harvard Dataverse.