



Text Mining and Natural Language Processing: Homework 2

Data Science Master's Program - BSE

Anastasiia Chernavskaia

Marvin Ernst

Viktoria Gagua

Contents

1	Introduction	1
2	Research Questions	1
3	Dictionary Creation	3
(a)	Data Collection	3
(b)	Text Preprocessing	3
(c)	TF-IDF Feature Extraction	4
4	United Nations General Debate Corpus	7
5	Results	10
(a)	SDG Representation in UN Speeches Over Time	10
(b)	Testing the Limitations of Our Dictionaries	14
6	Methodological Constraints and Future Considerations	17
7	Appendix	18

March 4, 2025

1 Introduction

Our goal is to identify whether specific topics—such as climate change, health, gender equality, and economic issues—have been addressed in United Nations General Assembly (UNGA) speeches over time. To ensure clarity and avoid misclassification, we construct topic-specific dictionaries based on descriptions of the 17 Sustainable Development Goals (SDGs).

To refine these dictionaries, we employ a TF-IDF (Term Frequency-Inverse Document Frequency) approach, which upweights terms that are distinctive to a given topic while down-weighting commonly used words that appear across multiple topics. This ensures that only terms strongly associated with a specific SDG are included in our analysis.

The use of SDG definitions as a reference framework is particularly relevant in the context of UNGA speeches, as these speeches tend to adopt formal language and cover broad policy issues that align closely with the SDGs. By leveraging this structured approach, we aim to track how different topics have been addressed over time and assess the extent to which major global events are reflected in UN discourse.

Additionally, we test the applicability of broader, less specific dictionaries, particularly for topics such as peace, justice, and strong institutions (SDG 16). Even though the expected keywords related to international conflicts such as "war", "military", and "terrorism" do not appear in the dictionary we created, one additional interest is to see whether such dictionary can still identify these conflicts based on the corpus of UNGA speeches.

2 Research Questions

1) SDG Representation in UN Speeches Over Time

For this part, we focus on the following questions:

a) Did the international movement on climate change find its way to the United Nations?

Climate change has become a central issue in global politics, with key agreements like the Kyoto Protocol, the Paris Agreement, and international climate strikes shaping discourse worldwide. This question examines whether and how these movements influenced UN General Assembly speeches, tracking the prominence of climate-related discussions (SDG 13) over time.

b) Are global milestones in Gender Equality agenda reflected in UNGA discourse?

Major turning points in the fight for gender equality—such as the Beijing Conference (1995), the founding of UN Women (2010), and the #MeToo movement (2017)—sparked international conversations and policy shifts. But did these moments translate into

stronger commitments and increased attention in the UN General Assembly? By analyzing speech trends over time, we explore whether global advocacy efforts and policy breakthroughs influenced the way world leaders addressed gender equality on the UN stage.

c) Do global health crises amplify the public voice in UN speeches?

From HIV/AIDS to Ebola and COVID-19, global health crises have repeatedly challenged international cooperation. But did these pandemics also shape discourse at the United Nations? By analyzing mentions of SDG 3 (Good Health and Well-being) over time, we investigate whether major health crises led to increased attention to public health in UN General Assembly speeches.

d) Echoes of the Economy: Do UN Speeches Reflect Global Economic Trends?

This question explores whether major economic events—such as recessions, financial crises, and shifts in global trade—are reflected in the discourse of UN General Assembly speeches. By analyzing SDG 8 (Decent Work and Economic Growth) mentions over time, we aim to assess whether global economic shifts influence the way economic topics are addressed at the UN.

2) Testing the Limitations of Our Dictionaries

In this section, we explore the boundaries of our approach by applying an SDG dictionary to a context where its effectiveness may be uncertain. Specifically, we examine whether a dictionary not explicitly designed for geopolitical discourse can still capture major historical events and distinctions between key international actors. For this, we ask the following research question:

Did NATO-aligned and Soviet-aligned countries differ in their discourse on Peace, Justice, and Strong Institutions (SDG 16) in UNGA speeches over time, and do observed trends align with historical expectations?

We investigate whether systematic differences exist in how NATO-aligned and Soviet-aligned countries addressed topics related to peace, justice, and institutional stability in their UNGA speeches throughout the Cold War and beyond. By applying our SDG 16 dictionary, we analyze the frequency of related terms across both groups. Given that SDG 16 is typically framed around governance, rule of law, and access to justice—rather than geopolitical conflict—we do not necessarily expect our dictionary to cleanly capture Cold War-era tensions. However, by pushing the limits of our methodology, we explore whether meaningful patterns can still emerge.

3 Dictionary Creation

(a) Data Collection

We use the UNDP SDG Index Corpus, which is available through the `datasets` library as `UNDP/sdgi-corpus` (as well as in UNDP's HuggingFace [repository](#)). This dataset contains a collection of documents related to the Sustainable Development Goals (SDGs), annotated with relevant metadata. The dataset is structured into two splits: a training set with 5,880 documents and a test set with 1,470 documents, resulting in a total of 7,350 documents.

Each document in the dataset includes several key features:

- **Text:** The raw textual content of the document.
- **Embedding:** A numerical representation of the document for machine learning applications.
- **Labels:** A sequence of integer values representing the SDGs associated with the document.
- **Metadata:** Additional information such as country, file ID, language, locality, document size, type, and year.

Since our analysis focuses on tracking SDG-related discourse over time in a standardized language, we filter the dataset to retain only English-language documents. After this filtering step, the dataset is reduced to 5,282 documents.

This corpus provides a diverse and structured collection of SDG-related texts, enabling a data-driven approach to understanding how different SDGs are discussed across various sources. By leveraging the metadata, we can further analyze patterns in SDG mentions across countries, document types, and time periods.

(b) Text Preprocessing

To ensure that the textual data is clean, standardized, and optimized for analysis, we apply a preprocessing pipeline to the corpus. The primary objectives of preprocessing are to remove noise, standardize the text format, and enhance the effectiveness of dictionary-based term matching.

First, we separate the dataset into individual dataframes for each SDG. This allows us to process the text while maintaining SDG-specific groupings. Each document undergoes a series of preprocessing steps, as outlined below.

Lowercasing and Cleaning

All text is converted to lowercase to ensure uniformity in term recognition. Additionally, we remove URLs, special characters, and numeric values, as these elements do not contribute meaningfully to our analysis.

Tokenization

Each document is split into individual tokens (words or phrases). We provide the option to use either the default NLTK tokenizer or spaCy's tokenization model. SpaCy is particularly beneficial for handling more complex linguistic structures.

Stopword Removal

Stopwords, which are common words that do not contribute significant meaning (e.g., *"the"*, *"is"*, *"and"*), are removed from the text. In addition to NLTK's standard English stopwords list, we define a custom list of domain-specific stopwords that frequently appear across multiple SDGs but do not contribute uniquely to their identification. This list includes terms such as *"sdg"*, *"policy"*, *"support"*, *"development"*, and *"government"*.

Lemmatization

Lemmatization reduces words to their base forms, ensuring that different inflections of the same word (e.g., *"running"* to *"run"*, *"countries"* to *"country"*) are treated as the same term. This process helps to standardize textual content and improves dictionary matching.

Batch Processing

Each SDG-specific dataframe undergoes this preprocessing pipeline in a batch process, this results in a clean and structured dataset.

(c) TF-IDF Feature Extraction

To identify the most important terms for each Sustainable Development Goal (SDG), we apply the Term Frequency-Inverse Document Frequency (TF-IDF) approach. TF-IDF is a widely used method for measuring the relevance of a term within a corpus by balancing term frequency with its uniqueness across documents. The TF-IDF score is computed as:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t)$$

where:

- $\text{TF}(t, d)$ is the term frequency, representing how often term t appears in document d .
- $\text{IDF}(t)$ is the inverse document frequency, which accounts for how common or rare a term is across all documents, given by:

$$\text{IDF}(t) = \log \frac{N}{n_t}$$

where N is the total number of documents and n_t is the number of documents containing term t .

Unlike standard document-based TF-IDF applications, we treat all text related to a single SDG as a single document. This allows us to extract key terms that best represent each SDG, ensuring that terms appearing frequently in one SDG category but rarely in others receive a higher TF-IDF weight.

Adjustments to Document Frequency

We experimented with `min df` and `max df` and discovered that very generous bounds (0.9 and 0.1) give us the most meaningful dictionaries.

Refining the Number of Terms and Weighting

Initially, we extracted the top 20 terms per SDG. Given the strong results, we expanded this to 50 terms while incorporating TF-IDF weights to improve performance. This ensures that less frequent but highly informative terms receive greater importance in our analysis.

Balancing Unigrams and Bigrams

One challenge in TF-IDF scoring is that both unigrams and bigrams (e.g., *"energy"* vs. *"renewable energy"*) often receive high scores. Keeping only unigrams removes critical contextual meaning, while keeping only bigrams may overlook the individual contributions of words appearing separately. To balance this:

- We retain both unigrams and bigrams.
- If a unigram is part of a frequently occurring bigram, its weight is slightly reduced to reflect its dependency on context.
- This allows for better distinction while ensuring that both individual words and meaningful phrases are captured.

Filtered Dictionary for SDG Classification

To refine our analysis further, we create a final dictionary by keeping only terms with high TF-IDF scores (above 0.1). This dictionary serves as the basis for later predictions, improving precision when applying the dictionaries to unseen speech data.

Applying Weights based on SDG Description TF

To further improve our dictionaries, we created 17 separate dictionaries on the basis of SDG descriptions from the [SDG website](#).

We created a dataframe with 17 rows, each row contained the SDG number and the corresponding textual description that was highly indicative of its respective SDG. For example, textual description of SDG 1 (No poverty) was very densely loaded with terms like *"poverty"*, *"extreme poverty"*, *"poor"*, *"vulnerable"*, etc. These texts helped us identify specific terms that are indicative of each SDG. After pre-processing, each SDG description was processed separately, and a term frequency (TF) dictionary was generated for each SDG, keeping only top 10 most frequent unigrams and bigrams.

Following the creation of the Description dictionaries, the next step was to compare them with the dictionaries created in the previous step using TF-IDF method. The goal of comparing these two sets of dictionaries was to upweight words that appear

in both (UNDP dataset and Descriptions dataset) while maintaining the original frequency for words that appear in only one of them. This was achieved by iterating through both dictionaries and adjusting word importance based on overlap. Words that appeared in both dictionaries received a higher weight factor (multiplying their frequency in the UNDP dataset by $(1 + \text{the relative frequency of the word in the Descriptions dataset})$), while words exclusive to one dataset retained their original frequency. This step ensured that the most relevant words received higher importance in the subsequent analysis.

SDG Dictionary Sample

Tables 1, 2 and 3 present a subset of the highest-ranked terms per SDG along with their TF-IDF scores, rounded to two decimal places.





SDG	Top Terms	SDG	Top Terms
	social protection (0.52) poverty line (0.32) poverty (0.31) poverty rate (0.27) extreme poverty (0.21) poor vulnerable (0.12)		stunting (0.33) malnutrition (0.32) sustainable agriculture (0.30) end hunger (0.28) food (0.25) obesity (0.24)
	death (0.32) health (0.27) live birth (0.24) mortality rate (0.19) hiv (0.19) maternal mortality (0.19)		quality education (0.36) education (0.29) enrollment (0.27) lifelong learning (0.25) secondary education (0.24) early childhood (0.22)
	violence (0.37) violence woman (0.21) woman (0.21) domestic violence (0.17) sexual (0.16) discrimination (0.15)		water (0.33) water quality (0.31) wastewater (0.30) basin (0.23) groundwater (0.23) management water (0.19)

Table 1: Top Terms Associated with Each SDG (1-6)







	<p>energy (0.39)</p> <p>energy consumption (0.35)</p> <p>energy efficiency (0.35)</p> <p>energy source (0.27)</p> <p>modern energy (0.25)</p> <p>clean energy (0.23)</p>		<p>unemployment rate (0.42)</p> <p>decent work (0.39)</p> <p>labour market (0.34)</p> <p>productive employment (0.20)</p> <p>employment decent (0.19)</p> <p>unemployed (0.18)</p>
	<p>resilient infrastructure (0.35)</p> <p>industrialization (0.33)</p> <p>passenger (0.29)</p> <p>foster innovation (0.28)</p> <p>broadband (0.27)</p> <p>infrastructure promote (0.24)</p>		<p>gini (0.35)</p> <p>discrimination (0.34)</p> <p>reduce inequality (0.27)</p> <p>social protection (0.25)</p> <p>inequality within (0.25)</p> <p>migrant (0.24)</p>
	<p>suwon (0.31)</p> <p>settlement inclusive (0.26)</p> <p>air quality (0.26)</p> <p>human settlement (0.25)</p> <p>city human (0.24)</p> <p>inclusive safe (0.23)</p>		<p>recycling (0.50)</p> <p>sustainable consumption (0.37)</p> <p>food waste (0.29)</p> <p>circular economy (0.29)</p> <p>plastic (0.27)</p> <p>sustainable (0.25)</p>

Table 2: Top Terms Associated with Each SDG (7-12)

The table illustrates how key terms align with their respective SDGs. For example, SDG 3 (Good Health and Well-being) is strongly associated with terms such as *"health"*, *"disease"*, and *"mortality"*, while SDG 7 (Affordable and Clean Energy) emphasizes terms like *"renewable energy"* and *"electricity"*.

By leveraging TF-IDF scoring, we ensure that our SDG dictionaries prioritize the most informative terms. These dictionaries are then applied to analyze UN General Debate speeches, allowing us to track SDG-related discourse over time.

4 United Nations General Debate Corpus

The dataset used in this study is the [United Nations General Debate Corpus](#) (UNGDC), which contains 10,760 official transcripts of speeches delivered at the United Nations General Assembly (UNGA) from 1946 to 2023. Each year, heads of state and high-






	<p>ghg emission (0.36)</p> <p>gas emission (0.32)</p> <p>climate (0.28)</p> <p>combat climate (0.21)</p> <p>paris agreement (0.20)</p> <p>change impact (0.19)</p>		<p>ocean (0.66)</p> <p>marine resource (0.38)</p> <p>fish (0.31)</p> <p>marine environment (0.26)</p> <p>protected area (0.25)</p> <p>marine (0.21)</p>
	<p>protected area (0.34)</p> <p>terrestrial ecosystem (0.23)</p> <p>land degradation (0.23)</p> <p>wildlife (0.23)</p> <p>soil (0.22)</p> <p>invasive (0.21)</p>		<p>corruption (0.36)</p> <p>violence (0.31)</p> <p>victim (0.24)</p> <p>court (0.21)</p> <p>anticorruption (0.19)</p> <p>criminal (0.18)</p>
	<p>debt (0.56)</p> <p>oda (0.41)</p> <p>global partnership (0.31)</p> <p>partnership sustainable (0.29)</p> <p>remittance (0.28)</p> <p>tax revenue (0.19)</p>		

Table 3: Top Terms Associated with Each SDG (13-17)

ranking representatives from all UN member states address the General Assembly, outlining their priorities and positions on key global issues.

Importantly, when a country is present, it delivers only one speech per year, making this dataset a unique and standardized record of international discourse. The corpus offers a valuable resource for analyzing how countries frame their foreign policy, respond to global events, and engage with long-term international challenges over time.

In addition to the speech texts, the dataset includes various metadata attributes that provide contextual information. These metadata fields include the country code, year, document ID, and session number. For a subset of the speeches, additional metadata such as democratic performance indicators, gender representation, major power status, and regime classification is also available. However, for 7,660 speeches, only the core metadata —country code, year, document ID, text, and session number— is present.

Despite these limitations, this core information is sufficient for our analysis, as our primary focus is on examining the textual content of the speeches over time and assessing how different countries have engaged with the SDGs in their discourse, rather than analyzing specific political classifications.

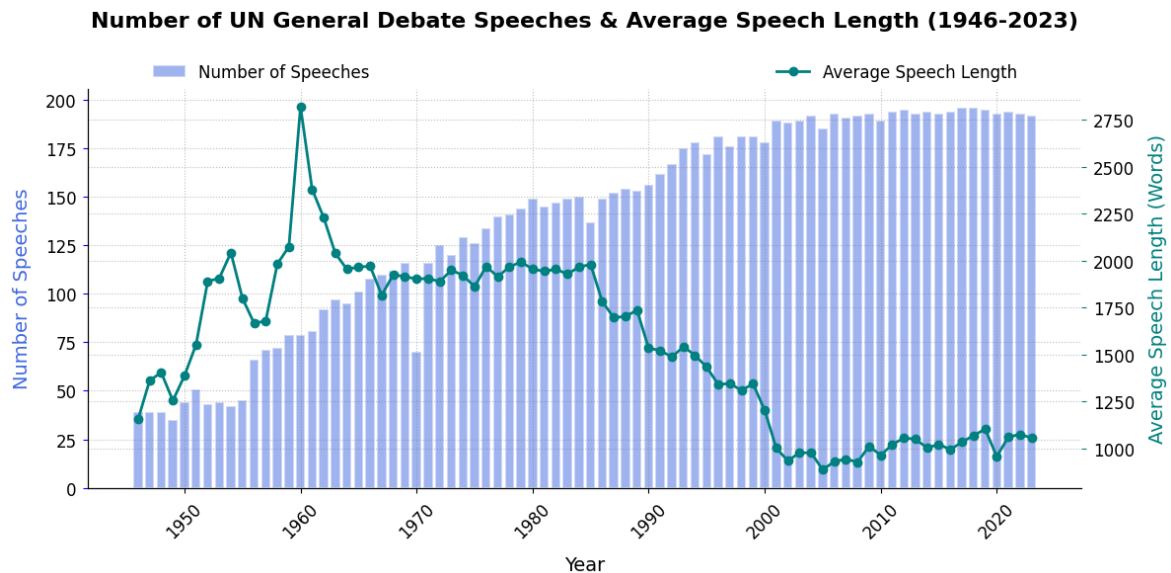


Figure 1: Number of speeches and average speech length over time

Judging from the Figure 1, the number of speeches in the UNGA General Debate has evolved alongside global political shifts. It grew steadily after 1946 as more nations joined the UN. During the Cold War (1960s-1980s), participation fluctuated due to geopolitical tensions, diplomatic boycotts, and major global events. Following the Cold War's end and the collapse of the Soviet Union (1991), speech numbers increased as new countries emerged. Global engagement also peaked after events like 9/11 (2001). Since 2021, participation has been slightly declining, probably influenced by COVID-19's virtual debates and evolving geopolitical tensions. Naturally, the average speech length has declined in the last few decades (from about 2000 to 1000 words per speech) to accommodate speakers from almost 200 Member States.

5 Results

(a) SDG Representation in UN Speeches Over Time

Did the international movement on climate change find its way to the United Nations?

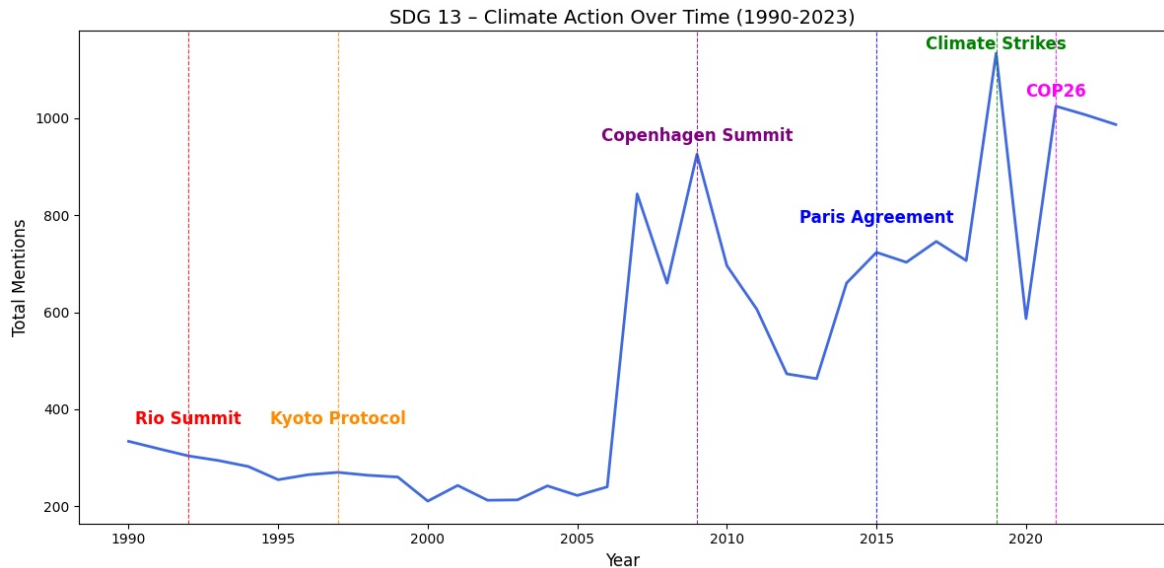


Figure 2: Temporal trends in the mentions of **SDG 13 (Climate Action)** in UN General Assembly speeches from 1990 to 2023. Key international climate events are highlighted along the x-axis. Note that term frequencies are weighted based on the SDG-specific dictionaries, which incorporate TF-IDF weightings to emphasize relevant terminology.

This graph shows how often people talked about climate action (SDG 13) from 1990 to 2023. From 1990 to 2005, climate action was not discussed much, even during the Rio Summit and Kyoto Protocol. Around 2005, mentions started growing rapidly, reaching a high point during the **2009 Copenhagen Summit**. After Copenhagen, interest dropped for a few years before rising again with the **2015 Paris Agreement**.

The biggest spike came in 2019 during the global climate strikes led by young activists like Greta Thunberg. In 2020, mentions dropped sharply (likely due to COVID-19) before recovering around **COP26 in 2021**. This thirty-year pattern shows how climate change has changed **from a niche topic to a major global concern**.

Importantly, our project work validates these findings, as the visible spikes in the graph clearly coincide with major summit years. This confirms that our methodology of developing specialized dictionaries and identifying relevant terminology was accurate and effective. The alignment between our data and known historical climate events shows how our analytical approach in capturing genuine patterns in climate discourse worked successfully.

Are global milestones in Gender Equality agenda reflected in UNGA discourse?

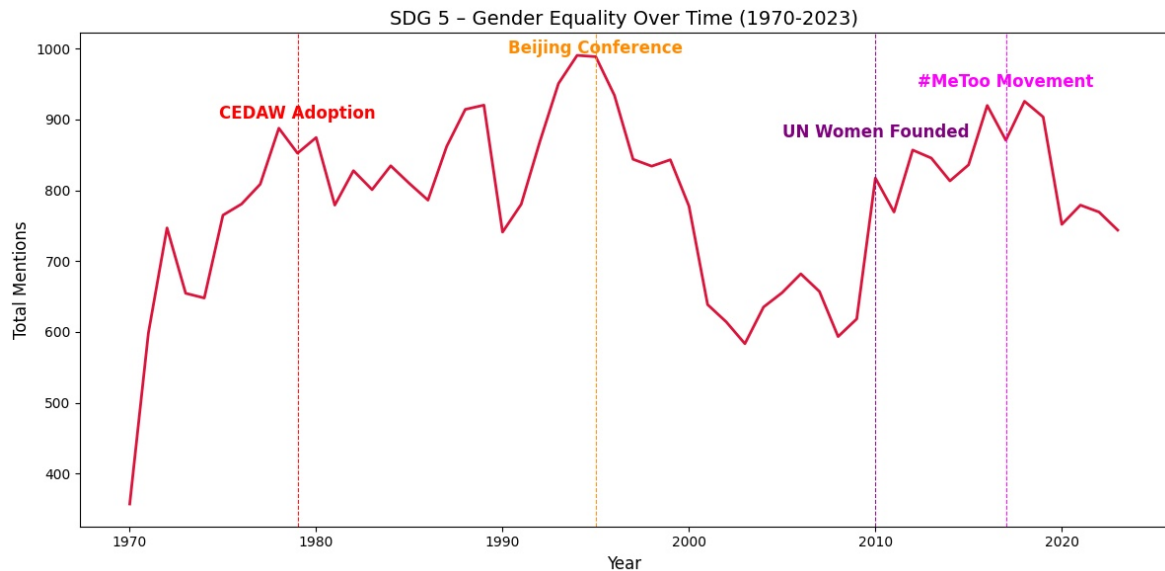


Figure 3: Temporal trends in the mentions of **SDG 5 (Gender Equality)** in UN General Assembly speeches from 1970 to 2023. Key international events are marked along the x-axis. Note that term frequencies are weighted based on the SDG-specific dictionaries, incorporating TF-IDF weightings.

This graph shows how often gender equality (SDG 5) was mentioned from 1970 to 2023. From 1970 to 1980, there was a sharp increase in mentions of gender equality, rising from around 350 to 900 weighted term appearances. The adoption of **CEDAW (Convention on the Elimination of All Forms of Discrimination Against Women) in 1979** coincides with a notable peak in the discussion.

The highest point on the graph appears during the **1995 Beijing Conference**, reaching about 1000 mentions. After Beijing, there was a significant decline in mentions for about a decade.

Around 2010, with the founding of **UN Women**, mentions began rising again. The **#MeToo movement** in 2017 corresponded with another major spike, showing how social media activism significantly impacted public discourse on gender equality.

Do global health crises amplify the public voice in UN speeches?

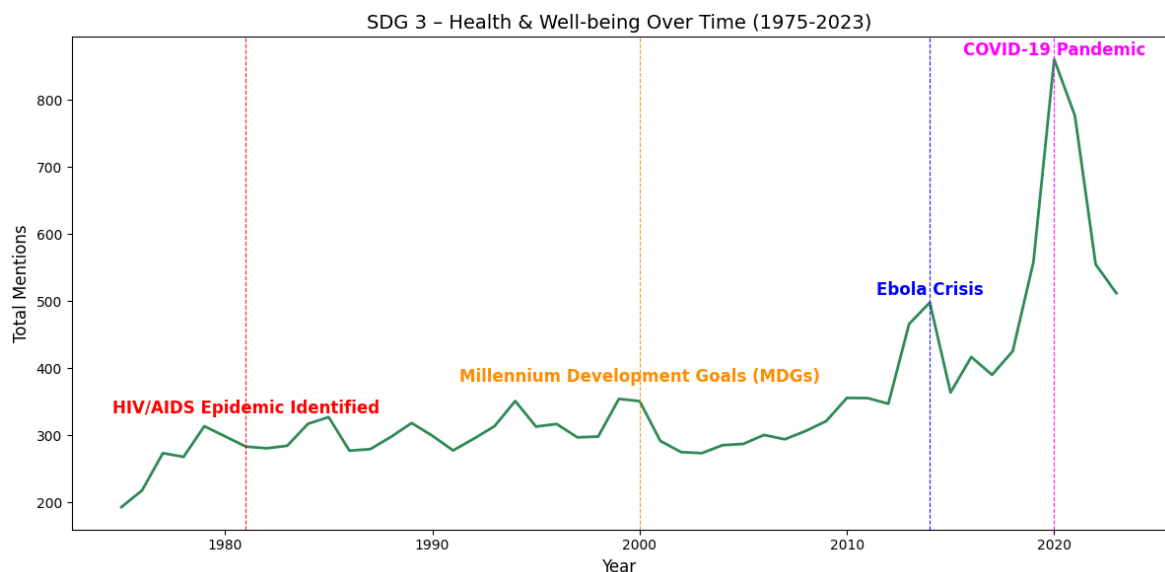


Figure 4: Temporal trends in the mentions of **SDG 3 (Good Health and Well-being)** in UN General Assembly speeches from 1975 to 2023. Key international health crises and policy milestones are marked along the x-axis. Term frequencies are weighted based on the SDG-specific dictionaries, incorporating TF-IDF weightings.

The trajectory of SDG 3 mentions in UN General Assembly (UNGA) speeches from 1975 to 2023 reflects the impact of major global health crises and policy initiatives. The first notable increase in health-related discussions aligns with the rise of the **HIV/AIDS epidemic** in the early 1980s, a period when the global response to the crisis intensified. However, despite the ongoing impact of HIV/AIDS, mentions of health topics remained relatively stable throughout the following decades.

A modest increase appears around the adoption of the **Millennium Development Goals (MDGs) in 2000**, which placed global health challenges—particularly infectious diseases and child mortality—at the center of international development efforts. However, it is only with the onset of acute health emergencies that we observe significant spikes in discourse.

The two most pronounced increases in mentions of SDG 3 correspond to two major global health crises: the **Ebola outbreak (2014)** and the **COVID-19 pandemic (2020–2021)**. The sharp rise in 2014 suggests that sudden outbreaks of infectious diseases can quickly elevate health issues in UN discussions. The most substantial spike occurred during the COVID-19 pandemic, as global health and well-being became dominant themes in international diplomacy and governance.

Echoes of the Economy: Do UN Speeches Reflect Global Economic Trends?

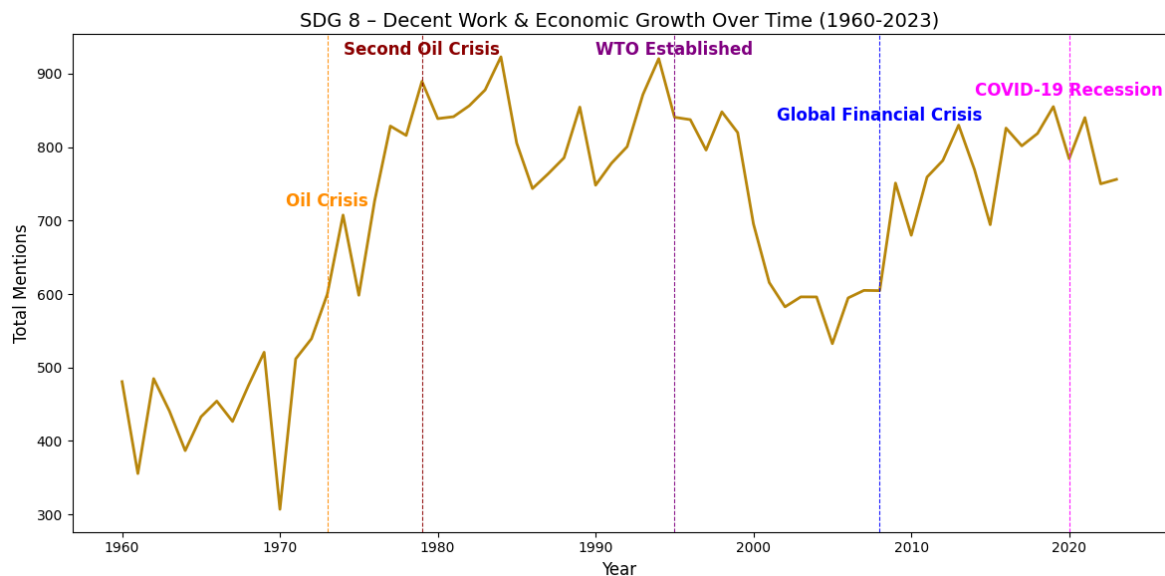


Figure 5: Temporal trends in the mentions of **SDG 8 (Decent Work and Economic Growth)** in UN General Assembly speeches from 1960 to 2023. Key international events are marked along the x-axis. Note that term frequencies are weighted based on the SDG-specific dictionaries, incorporating TF-IDF weightings.

The SDG 8 timeline (1960-2023) captures fluctuations in discussions related to economic growth and labor markets. The first major increase corresponds to the **1973 Oil Crisis**, which led to widespread economic instability. Mentions remained elevated through the **Second Oil Crisis in 1979**, highlighting how economic shocks influence UN discourse.

A notable increase occurs around the formation of the **World Trade Organization (WTO) in 1995**, reinforcing the global emphasis on economic liberalization. However, the largest spikes appear during financial crises, particularly the **2008 Global Financial Crisis** and the **2020 COVID-19 Recession**, which disrupted labor markets worldwide.

While the methodology effectively captures economic crises, SDG 8 covers broad themes such as employment, sustainable growth, and fair labor conditions. The increase in mentions does not necessarily imply a focus on workers' rights, but rather reflects how economic disruptions shape diplomatic discourse. Additionally, term weighting techniques may amplify certain topics over others, requiring caution in comparative interpretations.

(b) Testing the Limitations of Our Dictionaries

Did NATO-aligned and Soviet-aligned countries differ in their discourse on Peace, Justice, and Strong Institutions (SDG 16) in UNGA speeches over time, and do observed trends align with historical expectations?

In this analysis, we investigate whether systematic differences exist between NATO-aligned and Soviet-aligned countries in their discourse on SDG 16-related topics within UN General Assembly speeches over time. While NATO expanded over the years and the Soviet Union dissolved in 1991, for consistency, we maintain a fixed grouping of countries throughout the entire period of investigation.

The NATO-aligned countries included in this analysis are: *USA, Canada, United Kingdom, France, Germany, Italy, Netherlands, Belgium, Denmark, Norway, Greece, Turkey, Portugal, Spain, Luxembourg, and Iceland.*

The Soviet-aligned countries included are: *USSR, Poland, Hungary, Czechoslovakia, Bulgaria, Romania, Yugoslavia, Albania, East Germany (GDR), Cuba, Vietnam, China, and North Korea (PRK).*

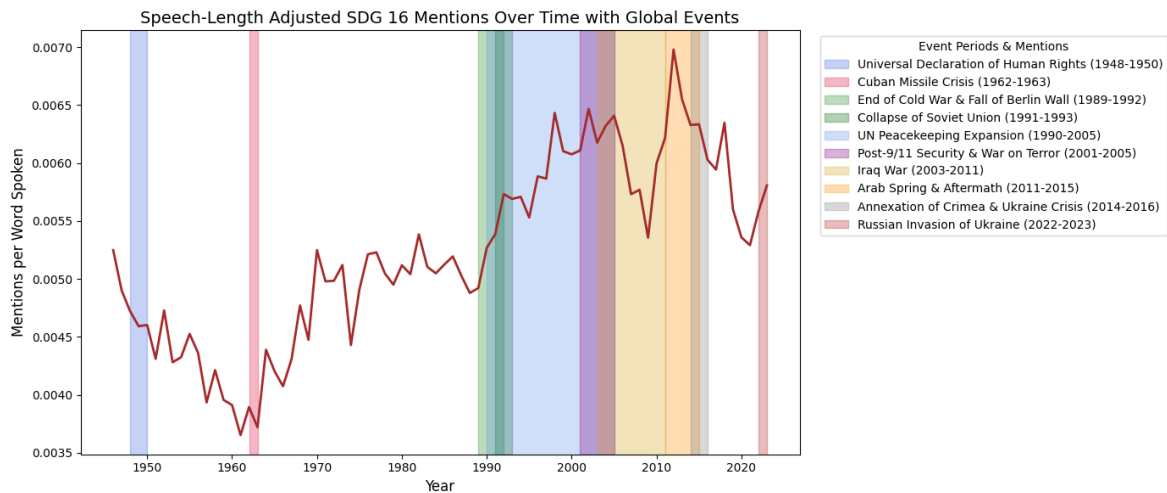


Figure 6: Temporal trends in **SDG 16 (Peace, Justice, and Strong Institutions)** mentions in UN General Assembly speeches, covering all countries from 1945 to 2023. The plot is adjusted by total words spoken that year. Key global events related to peace and conflict are highlighted along the x-axis. Term frequencies are weighted based on the SDG-specific dictionaries, which incorporate TF-IDF weightings.

This plot illustrates how SDG 16-related topics have been discussed in UN General Assembly speeches over time. While the dictionary was designed to capture terms associated with peace, justice, and institutions, it does not fully align with the terminology that would be most expected in discussions of major conflicts and wars. This highlights the limitations of dictionary-based approaches when dealing with broad and politically complex topics.

Nevertheless, certain historical events do stand out. For instance, we observe increases in SDG 16 mentions during the period of **UN Peacekeeping Expansion (1990–2005)**, the **9/11 attacks (2001)**, and a significant peak around the **Arab Spring (2011)**. These moments suggest that while our dictionary approach captures general trends in governance and institutional discourse, it may not fully capture the nuances of wartime rhetoric or security-focused discussions. This underscores the challenge of designing dictionaries that accurately reflect evolving political and diplomatic language in global institutions like the UN.

By comparing total mentions and speech-length adjusted mentions of SDG 16-related terms, we find that both groups largely follow similar patterns over time. However, identifying clear events directly associated with the Cold War or major conflicts where only one of the two blocs was involved proves challenging.

A notable peak in mentions by Soviet-aligned countries in **1960** was observed, which, upon further investigation, was driven primarily by increased references from Soviet representatives. Similarly, in the early **1970s**, NATO-aligned mentions rose sharply - this can be explained by the fact that additional NATO countries joined the UNGA, thus increasing the overall volume of speeches from NATO-aligned representatives.

Once adjusted for speech length, differences between the two groups become even less pronounced, reinforcing the idea that overall discourse on SDG 16-related topics followed similar long-term trends across both blocs. The increasing mentions in recent years suggest a growing focus on peace, justice, and institutional concerns within the UNGA.

There are several reasons why it may be difficult to directly link conflicts to speech content:

- **Dictionary Limitations:** The SDG 16 dictionary may not be perfectly suited for capturing the language used in discussions of specific conflicts.
- **Multiplicity of Conflicts:** Many different conflicts occurred throughout this period, involving various actors beyond just NATO- and Soviet-aligned countries.
- **Annual Speeches per Country:** Since each country delivers only one speech per year, large-scale analyses work well when considering all countries together but may be less reliable for identifying patterns in individual nations' speeches.

This analysis confirms our initial expectations: identifying war-related discourse using an SDG 16 dictionary designed for broader governance themes is highly challenging. However, we also reaffirm that dictionary-based approaches work best when the topic of interest is clearly and explicitly associated with a specific vocabulary. In contrast, when the connection is less direct—such as peace-related discourse overlapping with broader geopolitical rhetoric—human interpretation remains essential for fully understanding context and meaning.

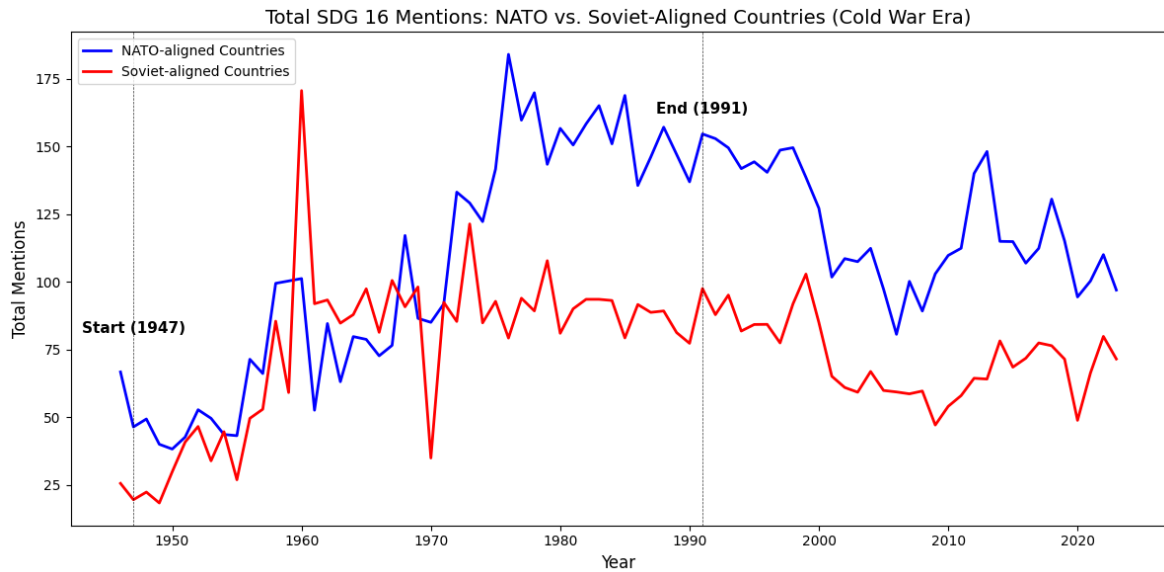


Figure 7: Comparison of **total mentions** of SDG 16-related terms in UN General Assembly speeches by NATO-aligned vs. Soviet-aligned countries from 1945 to 2023. The Cold War period (1947–1991) is marked to contextualize trends in discourse.

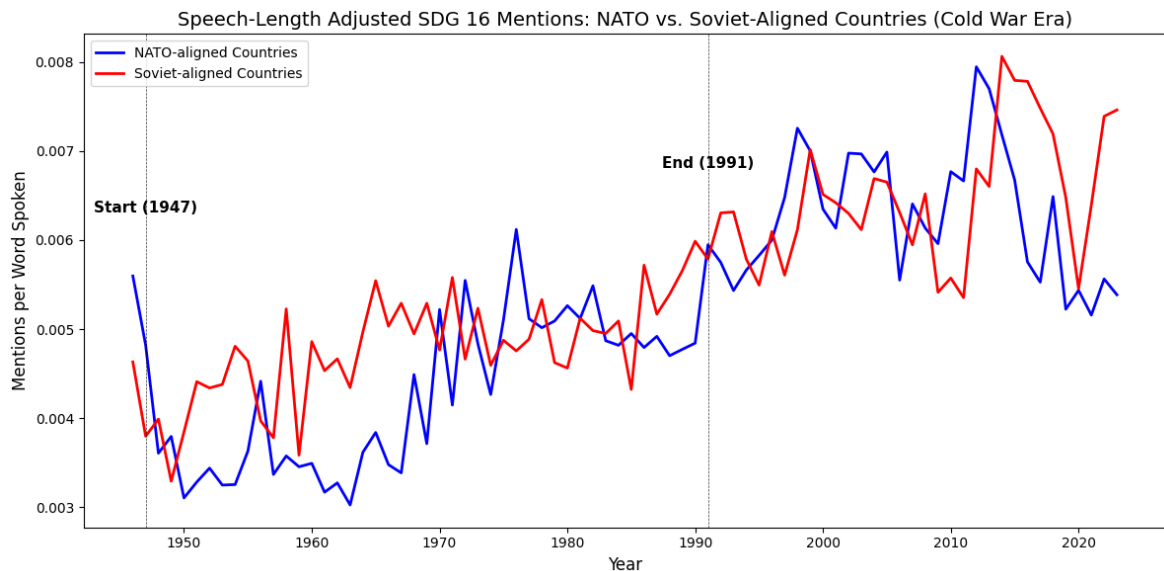


Figure 8: Speech-length adjusted trends in **SDG 16 mentions per word spoken** for NATO-aligned vs. Soviet-aligned countries from 1945 to 2023. The Cold War period (1947–1991) is highlighted to examine differences in discourse intensity.

6 Methodological Constraints and Future Considerations

Our project faces several important limitations that need further consideration to be improved:

Methodologically, our TF-IDF approach may not fully capture the nuanced context in which SDG-related terms appear. The dictionary-based method prioritized word frequency over semantic meaning, potentially missing implicit references or misinterpreting the topic that used different terminology to address similar concepts.

The UNGDC dataset itself presents limitations, as it only includes official speech transcripts and lacks informal diplomatic discussions where substantive policy coordination often occurs.

Our focus on English-speaking language speeches potentially introduces linguistic bias, as translated speeches may lose nuance or cultural context. Additionally, the varying length of speeches across different countries and years creates potential disparities in how we measure SDG mentions, potentially favoring nations that deliver longer speeches.

Finally, our analysis spans a period (1946-2023) where the SDGs did not exist for most of the timeframe, requiring us to retroactively apply contemporary frameworks to historical speeches. This creates kind of anachronic bias that may overinterpret past mentions of development issues as SDG-related.

Alternatively, instead of solely relying on TF-IDF, we could implement more sophisticated BERT (Bidirectional Representations from Transformers) model which helps understand the context and semantic meaning rather than just word frequency. For our SDG analysis, we could fine-tune BERT on SDG-related documents to recognize the relevant concepts even when they are expressed with different terminology. This would allow us to identify when a speech contains concepts like "renewable power infrastructure" as relevant to SDG 7, even without using exact dictionary terms. BERT's sentence embeddings could be used to measure similarity between speech segments and SDG definitions, providing more nuanced measurement than simple term counting.

Furthermore, we could implement LDA which is an unsupervised learning technique that would help us identify topics that naturally emerge from the speech corpus itself. In our context, LDA could discover how topics evolve over the decades of UN speeches, potentially revealing SDG-related themes before the goals were formally established. This would address the anachronism limitation by allowing historical speeches to define their own topic rather than retrofitting modern concepts.

7 Appendix

Limitations of SDG Comparisons

The following figures should not be interpreted as direct comparisons of Sustainable Development Goals (SDGs). Our SDG dictionaries were specifically designed to capture relevant topics within each SDG rather than to facilitate cross-SDG comparisons. Several methodological factors contribute to the lack of interpretability in these figures:

- **Dictionary Size Bias:** Each SDG dictionary contains a different number of terms, meaning that longer dictionaries naturally result in higher raw counts of mentions. This skews any direct comparison across SDGs.
- **Weighting Differences:** The terms within each dictionary were selected based on their TF-IDF scores, which reflect their relevance within specific SDG contexts. However, these weights are not normalized across SDGs, leading to inconsistencies in frequency measures.
- **Speech Length Variation:** The total length of speeches varies significantly across years and countries, influencing the absolute counts of SDG-related mentions. Some years may appear to emphasize certain SDGs more simply due to longer or more numerous speeches rather than genuine shifts in topic prominence.
- **Contextual Limitations:** While the dictionaries are effective at identifying speech segments related to an SDG, they do not capture the depth or framing of discussions. For instance, mentions of "climate" in SDG 13 do not necessarily reflect policy commitments or the same level of importance as mentions of "poverty" in SDG 1.
- **Differences in Political and Institutional Focus:** Certain SDGs (e.g., SDG 16 on peace and justice) may be inherently more relevant in UN debates than others (e.g., SDG 12 on responsible consumption). This institutional bias is not accounted for in raw mention counts.

For these reasons, the figures in this appendix serve only as supplementary material and should not be used for any direct comparative analysis between SDGs. A more rigorous approach, such as normalizing term frequencies relative to total speech length or adjusting for dictionary size differences, would be necessary to draw meaningful comparisons.

Figures

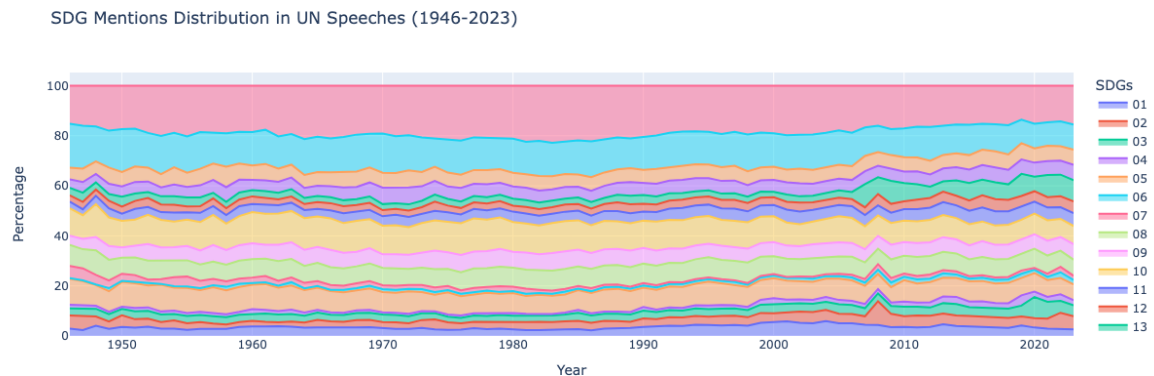


Figure 9: SDG Mentions Distribution in UN Speeches (1946-2003)

On this stacked area chart we see how mentions of each Sustainable Development Goal (SDG) contribute to the total SDG-related references in UN speeches over time. Each colored band represents one SDG, and the width of that band (from bottom to top) indicates the percentage share of references for that SDG in a given year.

For example, we see a smooth line for SDG 1 (bottom one), but there is a visible fluctuation throughout the whole period for SDG 12, depending on years it might spike or plummet rapidly;

Around 2007–2009, there is a subtle but noticeable shift in the balance of SDG references on the chart, which likely reflects the global economic turbulence of that period. Almost all SDGs are affected except SDG 1 and SDGs between 16-17;

Years 2019–2021 also show a uptick in references to health associated with SDG 3, but mostly all SDGs were affected as well. This change is consistent with the COVID-19 pandemic taking center stage, triggering heightened attention to public health, social protection, and broader systemic vulnerabilities.

The overall composition remains fairly steady, but those two periods—2007–2009 and 2019–2021—stand out as points where external global events reshaped the relative emphasis on different SDGs in UN speeches.

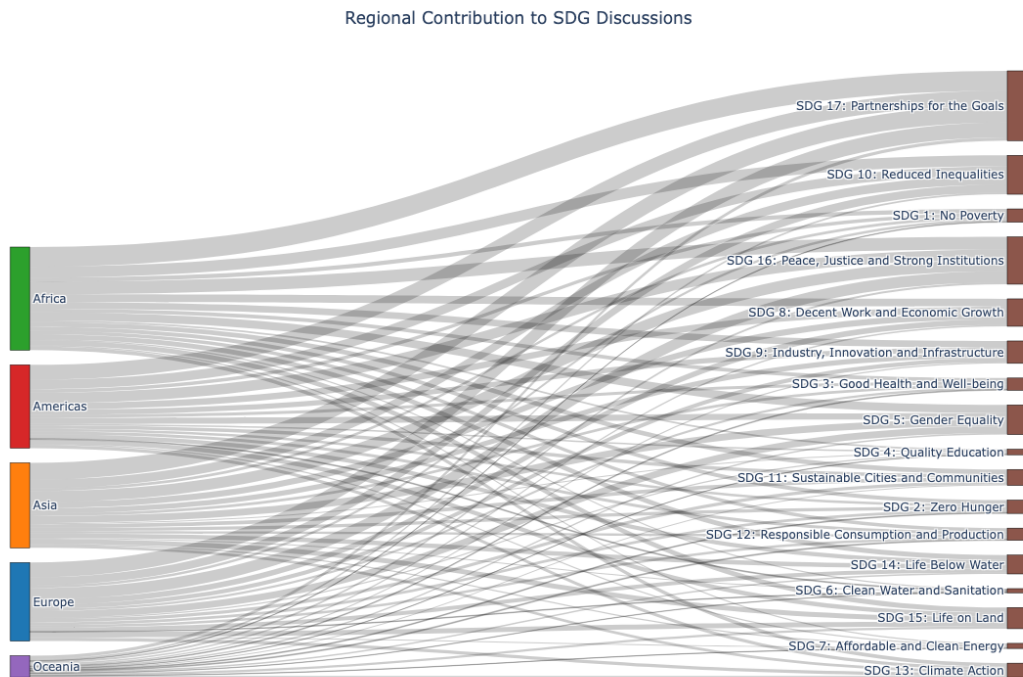


Figure 10: Given Sankey diagram depicts how frequently each region’s speeches reference each of the 17 Sustainable Development Goals.

Each flow or “stream” represents the extent to which a given region’s remarks align with a particular SDG. **Thicker lines generally mean stronger emphasis on that goal.** By tracing lines from left to right, we can see which goals get the most attention from each region and compare patterns of SDG priorities across different parts of the world.

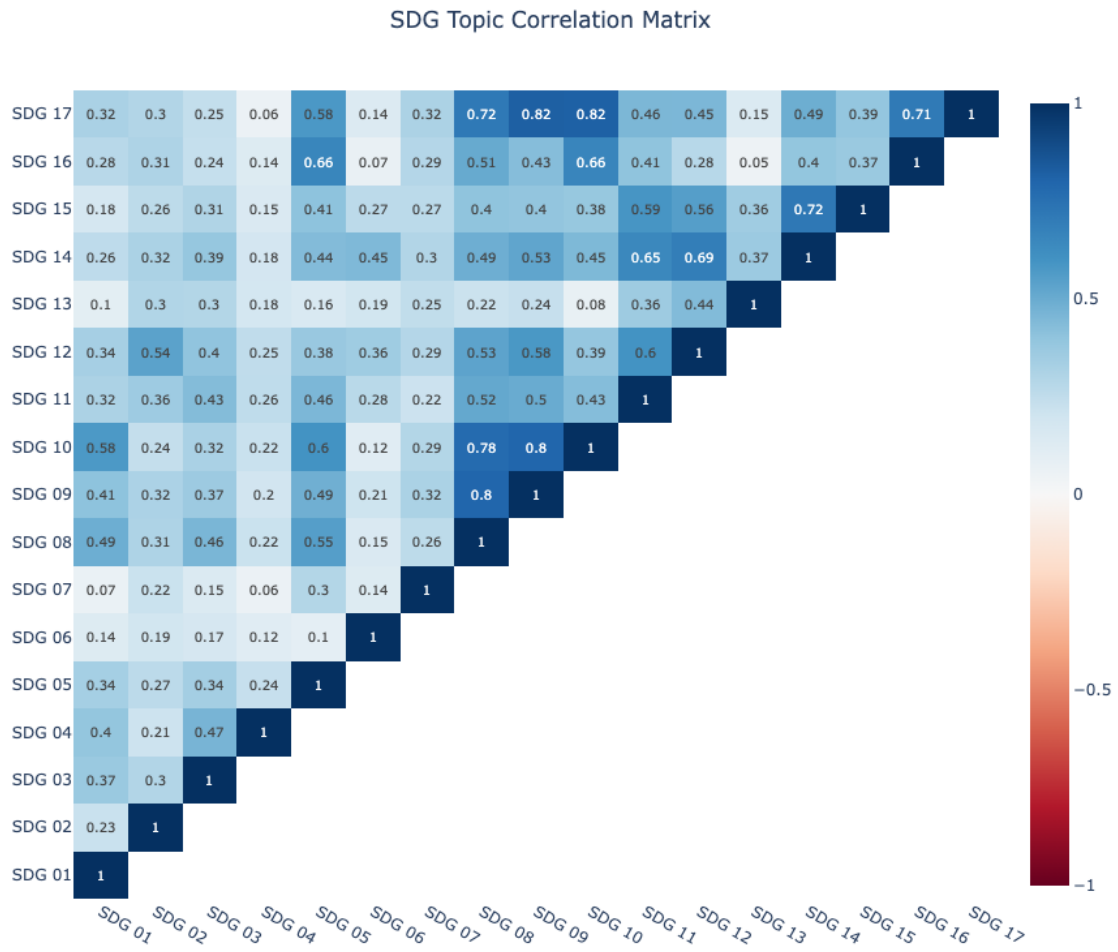


Figure 11: The Heatmap showing correlations between SDGs (which goals tend to be mentioned together)

It seems that SDG 09 and SDG 17 are most correlated with 0.82 as well as SDG 17 and SDG 10; We see the high correlation between SDG 9 (Industry, Innovation, and Infrastructure) and SDG 17 (Partnerships for the Goals) because infrastructure development often hinges on international collaboration, technology transfer, and capacity-building partnerships. Speeches referencing new infrastructure projects or innovations frequently also cite the need for external funding, shared research, or cross-border alliances, which fall under SDG 17.

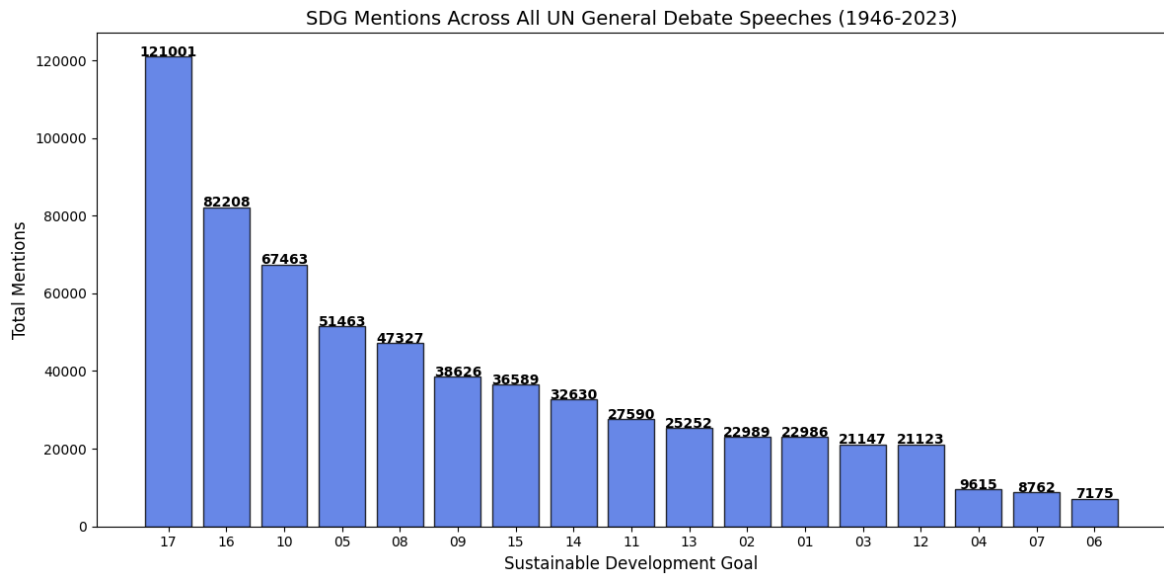


Figure 12: SDG 17 (Partnerships for the Goals) shows the most total mentions in UN General Debate speeches—suggesting that countries often talk about the need for global collaboration, financing, and technology transfer when they discuss sustainable development. SDG 16 (Peace, Justice, and Strong Institutions) and SDG 10 (Reduced Inequalities) also rank very high, likely because terms related to governance, conflict, and inequality surface frequently in diplomatic statements. Meanwhile, goals focused on areas such as water and sanitation (SDG 6) and affordable/clean energy (SDG 7) appear toward the lower end of the chart, which might reflect that countries reference them less explicitly (even if they are important priorities). Or it might be that the vocabulary around them is narrower in formal speeches.

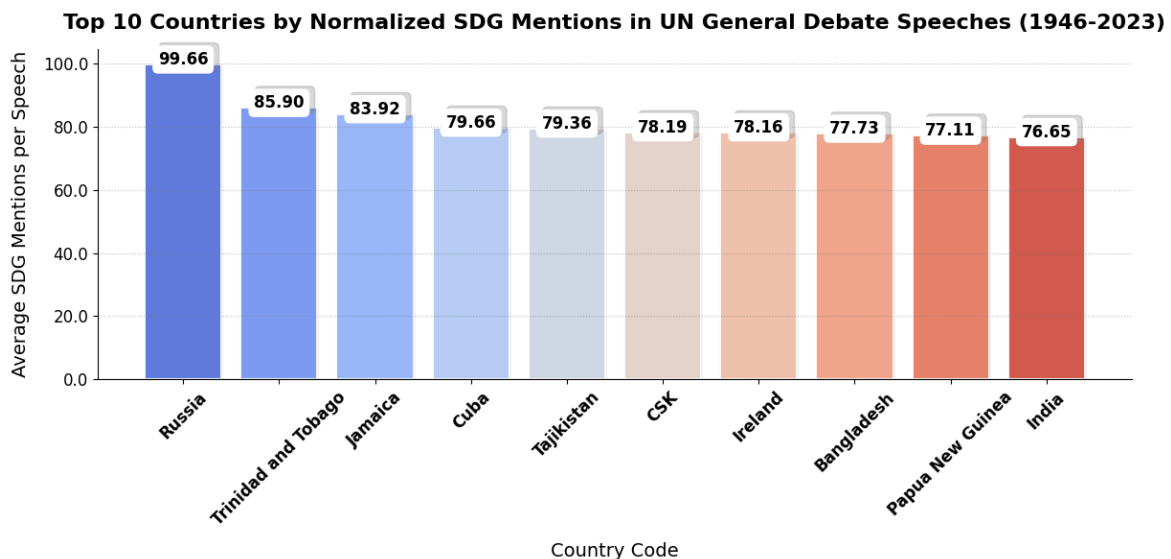


Figure 13: Top 10 Countries by Normalized SDG Mentions in UN General Debate Speeches (1946-2023)

Figure 13 shows the top 10 countries with the highest average number of SDG mentions per speech in the UNGDC. The values represent the average number of SDG-related mentions per speech for each country.

- Russia (RUS) leads with an average of 99.66 mentions per speech, meaning that SDG-related topics are highly emphasized in Russian speeches.
- Trinidad and Tobago (TTO) and Jamaica (JAM) follow closely, suggesting that even smaller countries can have a strong focus on SDGs.
- Cuba (CUB) and Tajikistan (TJK) also rank high, indicating a strong presence of SDG themes in their discourse.
- The list includes a mix of large economies (e.g., India, Russia, Ireland) and smaller nations (e.g., Papua New Guinea, Bangladesh, Jamaica, Trinidad and Tobago), showing that SDG emphasis is not strictly correlated with economic size or geopolitical power.
- Some expected major players (e.g., the U.S., China, Germany) are absent, which might suggest they either have more diluted SDG mentions or focus on other policy areas in their speeches.

Full SDG Dictionaries

SDG	Name	Top Terms (TF-IDF Score)
(1)	No Poverty	social protection (0.52), poverty line (0.32), poverty (0.31), poverty rate (0.27), poverty form (0.25), end poverty (0.22), extreme poverty (0.21), pension (0.18), living poverty (0.16), social assistance (0.16), risk poverty (0.15), elderly (0.15), everywhere (0.14), social service (0.13), poor vulnerable (0.13), basic service (0.12), social exclusion (0.11), deprivation (0.11), resource (0.11)
(2)	Zero Hunger	stunting (0.33), malnutrition (0.32), sustainable agriculture (0.30), end hunger (0.28), food (0.26), obesity (0.25), food insecurity (0.22), overweight (0.22), food production (0.20), nutritional (0.19), genetic resource (0.15), vegetable (0.15), agricultural (0.15), agricultural production (0.14), nutritious (0.14), zero hunger (0.14), meal (0.13), seed (0.13), ane (0.13), soil (0.13), food nutrition (0.12), diet (0.12), including (0.11), agricultural land (0.11), market (0.10), hunger (0.10)
(3)	Good Health and Well-being	death (0.32), health (0.27), live birth (0.24), mortality rate (0.19), hiv (0.19), maternal mortality (0.19), health care (0.16), mental health (0.16), noncommunicable disease (0.16), tuberculosis (0.16), birth (0.16), disease (0.15), tobacco (0.15), neonatal (0.14), patient (0.14), health-care (0.14), hepatitis (0.13), vaccine (0.13), malaria (0.13), maternal (0.12), drug (0.11), ncds (0.11), wellbeing age (0.11), cancer (0.11), suicide (0.10), mortality (0.10), medicine (0.10)
(4)	Quality Education	quality education (0.36), education (0.29), enrolment (0.27), lifelong learning (0.25), secondary education (0.24), early childhood (0.22), literacy (0.21), promote lifelong (0.17), equitable quality (0.16), learning opportunity (0.16), education promote (0.16), tvet (0.14), preprimary (0.14), language (0.14), primary education (0.13), learner (0.13), inclusive equitable (0.13), mathematics (0.13), childhood education (0.12), vocational education (0.12), educaton (0.12), reading (0.12), ensure inclusive (0.12), parent (0.11), lifelong (0.11), ensure (0.11), technical vocational (0.11), inclusive education (0.10)
(5)	Gender Equality	violence (0.37), violence woman (0.21), woman (0.21), domestic violence (0.18), sexual (0.16), discrimination (0.15), marriage (0.14), victim (0.13), woman men (0.12), achieve gender (0.12), proportion woman (0.12), genderbased violence (0.11), sexual violence (0.11), equal (0.10), gender (0.10)
(6)	Clean Water and Sanitation	water (0.33), water quality (0.31), wastewater (0.30), basin (0.23), groundwater (0.23), management water (0.20), wastewater treatment (0.18), sanitation service (0.18), treatment plant (0.16), sanitation (0.16), sustainable management (0.15), water body (0.15), surface water (0.15), water source (0.14), sewage (0.14), freshwater (0.13), toilet (0.13), wash (0.13), ane (0.13), water use (0.13), sewerage (0.13), irrigation (0.12), suwon (0.11)
(7)	Affordable and Clean Energy	energy (0.38), renewable energy (0.34), electricity (0.32), access energy (0.28), electricity access (0.25), clean energy (0.23), fossil fuel (0.22), sustainable energy (0.21), energy efficiency (0.20), fuel (0.19), modern energy (0.18), consumption (0.17), efficiency (0.16), solar (0.15), household energy (0.14), bioenergy (0.14), electricity generation (0.14), sustainable (0.13), electrification (0.13), energy source (0.12), gas (0.12), including (0.11), power (0.11), energy transition (0.10)

SDG	Name	Top Terms (TF-IDF Score)
(8)	Decent Work and Economic Growth	employment (0.35), economic growth (0.33), work (0.31), productivity (0.28), unemployment (0.26), informal employment (0.24), labor market (0.23), worker (0.22), decent work (0.21), industry (0.20), financial (0.19), wages (0.18), sustainable economic (0.17), child labor (0.16), gdp (0.15), labor (0.15), sector (0.14), enterprise (0.14), growth (0.13), youth unemployment (0.13), labor force (0.12), innovation (0.12), entrepreneurship (0.11), productivity growth (0.11), economy (0.10), tourism (0.10)
(9)	Industry, Innovation and Infrastructure	infrastructure (0.37), industry (0.35), innovation (0.32), industrialization (0.30), technology (0.28), research development (0.25), r and d (0.23), transport infrastructure (0.22), resilient infrastructure (0.21), gdp industry (0.20), internet (0.19), connectivity (0.18), manufacturing (0.18), rail (0.17), engineering (0.16), digital infrastructure (0.15), road transport (0.15), mobile broadband (0.14), investment (0.13), financial (0.13), patent (0.12), capacity (0.12), logistic (0.11), airport (0.11), urban infrastructure (0.10), technology transfer (0.10)
(10)	Reduced Inequalities	inequality (0.36), income (0.34), social protection (0.32), discrimination (0.30), economic inequality (0.28), inclusion (0.26), wealth (0.24), migrant (0.22), redistribution (0.21), policy (0.20), tax (0.19), wage gap (0.18), wealth distribution (0.17), accessibility (0.16), social mobility (0.16), gini coefficient (0.15), vulnerable (0.14), disparity (0.14), marginalized (0.13), ensuring equality (0.12), globalization (0.12), progress (0.11), socioeconomic status (0.11), universal (0.10)
(11)	Sustainable Cities and Communities	urban (0.39), city (0.37), urban planning (0.34), transport (0.32), housing (0.30), air pollution (0.28), urbanization (0.27), smart city (0.25), slum (0.24), infrastructure (0.22), sustainable transport (0.21), resilience (0.20), public transport (0.19), affordable housing (0.18), sustainable city (0.17), population density (0.16), accessibility (0.15), urban sprawl (0.14), climate resilience (0.14), sustainable (0.13), metropolitan (0.12), flood risk (0.12), zoning (0.11), road safety (0.11), waste management (0.10)
(12)	Responsible Consumption and Production	sustainable consumption (0.38), production (0.36), waste (0.34), recycling (0.31), supply chain (0.30), resource efficiency (0.28), circular economy (0.26), sustainable production (0.25), material footprint (0.23), plastic waste (0.22), food waste (0.20), consumption (0.19), sustainable procurement (0.18), corporate responsibility (0.17), ecolabel (0.16), lifecycle (0.16), sustainable lifestyle (0.15), reuse (0.14), waste management (0.13), responsible consumption (0.12), chemical waste (0.12), consumer awareness (0.11), industrial waste (0.11), ensuring sustainability (0.10)
(13)	Climate Action	climate change (0.40), greenhouse gas (0.38), carbon emissions (0.36), mitigation (0.34), adaptation (0.32), temperature rise (0.30), renewable energy (0.28), fossil fuel (0.26), climate policy (0.24), extreme weather (0.22), carbon neutrality (0.21), paris agreement (0.20), resilience (0.19), sustainability (0.18), emission reduction (0.16), deforestation (0.15), sea level rise (0.15), natural disaster (0.14), co2 emissions (0.13), sustainable development (0.13), energy efficiency (0.12), biodiversity loss (0.11), afforestation (0.10)
(14)	Life Below Water	ocean (0.39), marine biodiversity (0.37), fishery (0.35), pollution (0.32), plastic pollution (0.30), coral reef (0.28), overfishing (0.26), sustainable fishery (0.24), marine ecosystem (0.22), ocean acidification (0.20), sustainable seafood (0.19), water pollution (0.18), conservation (0.17), marine resources (0.16), fish stock (0.15), deep sea mining (0.14), coastal erosion (0.13), maritime (0.12), sustainable management (0.12), biodiversity conservation (0.11), marine protected area (0.10)

SDG	Name	Top Terms (TF-IDF Score)
(15)	Life on Land	biodiversity (0.41), forest (0.39), land degradation (0.37), desertification (0.35), reforestation (0.32), deforestation (0.30), ecosystem (0.28), conservation (0.26), sustainable forestry (0.24), species loss (0.22), protected area (0.21), soil degradation (0.20), afforestation (0.19), wildlife (0.18), restoration (0.16), ecosystem services (0.15), ensuring sustainability (0.14), ecological balance (0.14), sustainable management (0.13), ensuring (0.12), green corridors (0.11), environmental protection (0.11), forest cover (0.10)
(16)	Peace, Justice and Strong Institutions	corruption (0.36), violence (0.31), victim (0.24), court (0.21), anticorruption (0.19), criminal (0.18), institution (0.15), birth (0.14), legal aid (0.14), ensure (0.13), judicial (0.13), peace (0.13), rule law (0.12), institution level (0.12), accountable (0.12), justice (0.12), peaceful inclusive (0.11), abuse (0.11), peace justice (0.11), strong institution (0.10), form violence (0.10), strengthen (0.10), combat (0.10), crime (0.10)
(17)	Partnerships for the Goals	debt (0.56), oda (0.41), global partnership (0.31), partnership sustainable (0.29), remittance (0.28), mean implementation (0.23), tax revenue (0.20), broadband (0.17), wto (0.16), technology (0.15), sustainable (0.15), including (0.13), partnership (0.11), peace (0.11), agadir (0.11), bad köstritz (0.10), lesotho (0.10), goal (0.10)

Table 4: Top Terms with TF-IDF Scores for Each SDG (threshold score 0.1)