Indian Institute of Technology (Indian School of Mines), Dhanbad

(Department of Humanities and Social Sciences)

Project Title:

"Revealing the Predominant Topic in WHO Speeches (2000–2024): An Investigation Based on Topic Modeling and Clustering "

Submitted by:

Parul Priya
Roll No.: [24MA0013]

Submitted to:

[Dr. ShanmugapriyaT]
Department of Humanities and Social Sciences
IIT (ISM), Dhanbad

📅 **Academic Year: 2024–2026**

Title: Revealing the Predominant Topic in WHO Speeches (2000–2024): An Investigation Based on Topic Modeling and Clustering

Introduction

Since its founding, the World Health Organization (WHO) has been essential in directing international health policy, organizing crisis responses, and offering strategic and scientific leadership during serious medical catastrophes. Its speeches are a rich textual corpus that reflects its alliances, policy viewpoints, and priorities in addition to being public communications. Significant public health events, such as the HIV/AIDS pandemic, SARS, H1N1, Ebola, and most importantly, COVID-19, occurred worldwide between 2000 and 2024. WHO had a vital voice during these crises, influencing global health responses and highlighting the need for international collaboration. In addition to being reactive, these statements have a significant role in establishing global health agendas and emphasizing preparedness, equity, and sustainable development.

Using text mining techniques, this study examines these speeches from a data-driven standpoint in order to identify recurrent themes and the most often discussed subjects. The work uses Named Entity Recognition (NER), Topic Modeling (LDA), Clustering (K-Means), and Frequent Word Analysis in Natural Language Processing (NLP) to glean valuable information from WHO speeches' textual content. When combined, these techniques help us go beyond simple word frequency and reveal the complex patterns of WHO's health communication across time.

We were able to uncover WHO's operational scope and response areas by using NER to identify important locations, organizations, and diseases that are commonly addressed, such as "Gaza," "COVID-19," and "UNICEF." We were able to identify the most often occurring terms through frequent word analysis, which provided a basic understanding of the corpus's emphasis. Health emergencies, maternal health, humanitarian crises, noncommunicable illnesses, and the sustainability of the health system were the five main themes identified via topic modeling. While PCA-enabled word clustering graphically illustrated conceptual connections between talks, clustering techniques assisted in the classification of utterances into logical topic categories.

What was the most discussed issue in WHO speeches between 2000 and 2024, according to this paper? The goal is to thoroughly examine how WHO's communications relate to global health events and policy priorities, rather than just listing prevalent words and phrases. Health communicators, legislators, and academics with an interest in digital humanities, data-driven decision-making, and global health governance will find this research to be contemporary and pertinent. In the end, the study emphasizes how text mining may be used to comprehend institutional narratives, the development of policies, and the discursive framing of global health issues.

Review of Literature

Because it provides automatic and reproducible insights into massive textual datasets, text mining has emerged as a crucial tool in health communication research (Murphy et al., 2021). Natural Language Processing (NLP) approaches have being used more and more to identify patterns, identify themes, and expose discursive priorities as digital corpora of speeches, publications, and reports become more widely available. Blei et al. (2003) presented Latent Dirichlet Allocation (LDA), a fundamental technique for identifying abstract topics in sizable document collections, while Camacho-Collados and Pilehvar (2018) highlighted the importance of word and sense embeddings in comprehending contextual meanings in health-related text. Such methods have been used in recent studies to investigate media narratives, public health messages, and political discourse. The usefulness of subject modeling in political speeches was illustrated by Olovsson and Öberg (2020), who showed how political priorities change over time. To find important behavioral factors, Mehrolia et al. (2021) also looked at public health messages during the COVID-19 epidemic. According to these studies, machine learning techniques not only identify reoccurring topics but also aid in understanding how organizations formulate problems over time.

Few studies, nevertheless, have thoroughly examined WHO's remarks using a variety of techniques. By integrating NER, topic modeling, clustering, and frequent word analysis, this study seeks to close this gap and provide a solution to the main question: Which topic has been discussed the most in WHO speeches over the last 20 years?

Research Question

What is the most explored topic in WHO speeches between 2000 and 2024?

Objectives

1. **To find keywords and recurrent themes in WHO talks over time:**To find recurring themes and changing health priorities in WHO's public communications, this entails examining the frequency of important words throughout the corpus.

2. **To use NER to extract important named entities (such as nations, illnesses, and organizations):** In order to draw attention to WHO's operational emphasis and areas of concern in the context of global health discourse, Named Entity Recognition (NER) recognizes and categorizes proper nouns such as illnesses, places, and institutions.

3. **To use topic modeling to identify abstract themes and their development:** Uncovering hidden thematic patterns in the text by topic modeling, particularly with LDA,

demonstrates how subjects like pandemics or nutrition evolve and change in importance over time.

4. **To arrange related thematic speeches for trend analysis using clustering:**
   Researchers can observe how many speeches align around central concerns and how discourse shifts during significant occurrences like pandemics by clustering speeches based on text similarity.

5. **To present findings with visualizations such as bar charts, word clouds, pie charts, and heatmaps:**
   Visual tools make complex data understandable by highlighting word frequencies, topic distributions, and co-occurrence patterns, helping audiences grasp major trends in WHO communication.

6. **To interpret results in alignment with WHO's known health priorities:**
   Analyzing the data in light of WHO's strategic plans ensures that identified topics accurately reflect institutional objectives like universal health coverage, emergency preparedness, and health equity.

7. **To discuss how these methods help the audience understand global health communication:**
   Text mining techniques provide structured insights from large texts, making it easier for researchers and policymakers to understand how WHO shapes public narratives and influences health policy.

Methodology

The dataset comprises speeches from WHO between 2000 and 2024, primarily consisting of transcriptions of key addresses, media briefings, and public health updates. The methods employed are detailed below:

- **Named Entity Recognition (NER):**
  NER is a Natural Language Processing technique used to identify and classify specific entities in text, such as countries, diseases, organizations, and key individuals. In this study, spaCy was used to implement NER, allowing us to extract structured information from unstructured speech texts. By detecting entities like "COVID-19," "UNICEF," or "Gaza," the method helps map WHO's primary areas of concern, operational regions, and key collaborators, highlighting how these entities are represented across different contexts and time periods.

- **Frequent Word Analysis:**
  This technique identifies the most commonly used words in a corpus. Implemented using NLTK, it involves removing stop words (e.g., “and,” “the”) and calculating the frequency of meaningful words. This helps reveal recurring ideas and central themes in WHO speeches. Terms such as “health,” “support,” and “pandemic” emerged frequently, indicating WHO’s focus on global health emergencies and cooperation. Frequent word analysis serves as a foundational method for understanding the dominant language patterns and framing in institutional communication.

- **Topic Modeling (LDA):**
  Latent Dirichlet Allocation (LDA) is a statistical model that uncovers hidden thematic structures within a collection of documents. It assumes each document contains a mixture of topics, and each topic is characterized by a distribution of words. In this study, LDA was applied to the WHO speech corpus, and coherence scores were used to select five meaningful topics. Each topic was interpreted based on the most relevant words, providing insights into key areas such as health emergencies, maternal care, and funding issues.

- **Clustering (K-means):**
  K-means is an unsupervised machine learning algorithm used to group data into clusters based on similarity. Here, Term Frequency-Inverse Document Frequency (TF-IDF) vectors of the speeches were created to capture their unique word features. These vectors were then clustered using K-means to group speeches with similar themes. Silhouette scores were used to determine the quality of clusters. This technique helped categorize speeches into coherent thematic groups, such as pandemic responses or humanitarian aid, enabling a clearer understanding of WHO’s messaging patterns.

- **Principal Component Analysis (PCA) for Word Clusters:**
  PCA is a dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space while retaining the most important features. In this study, PCA was used to reduce the dimensionality of word vectors generated from the corpus. The transformed data was then clustered to visualize groups of semantically similar words. This helped identify conceptual groupings, such as health-related terms or crisis-related terms, and provided a visual summary of the thematic structure of WHO’s discourse, enhancing interpretability and comparison of key concepts.

Analysis

Frequent Word Analysis:
The most frequent terms include "health," "emergency," "support," "countries," "system," "funding," and "pandemic." The prevalence of these words indicates WHO’s consistent focus on maintaining global health security and responding to outbreaks. Other notable mentions include “maternal,” “nutrition,” “vaccines,” and “Gaza,” signifying a secondary emphasis on vulnerable populations and geopolitical crises.

NER Findings:
NER identified several prominent entities: geographic locations like "Gaza," "Myanmar," "South Africa," "Germany"; diseases like "COVID-19," "cholera," "malaria"; organizations such as "UNICEF," "USAID"; and terms like "funding," "ceasefire," and "emergency." The regular mention of these entities reveals WHO's operational domains and collaborative network. For example, repeated references to "Gaza" point to WHO's focus on conflict-affected regions. Similarly, the frequent citation of vaccine types and health initiatives highlights key public health campaigns.

Topic Modeling Results:
The LDA model extracted five coherent topics:

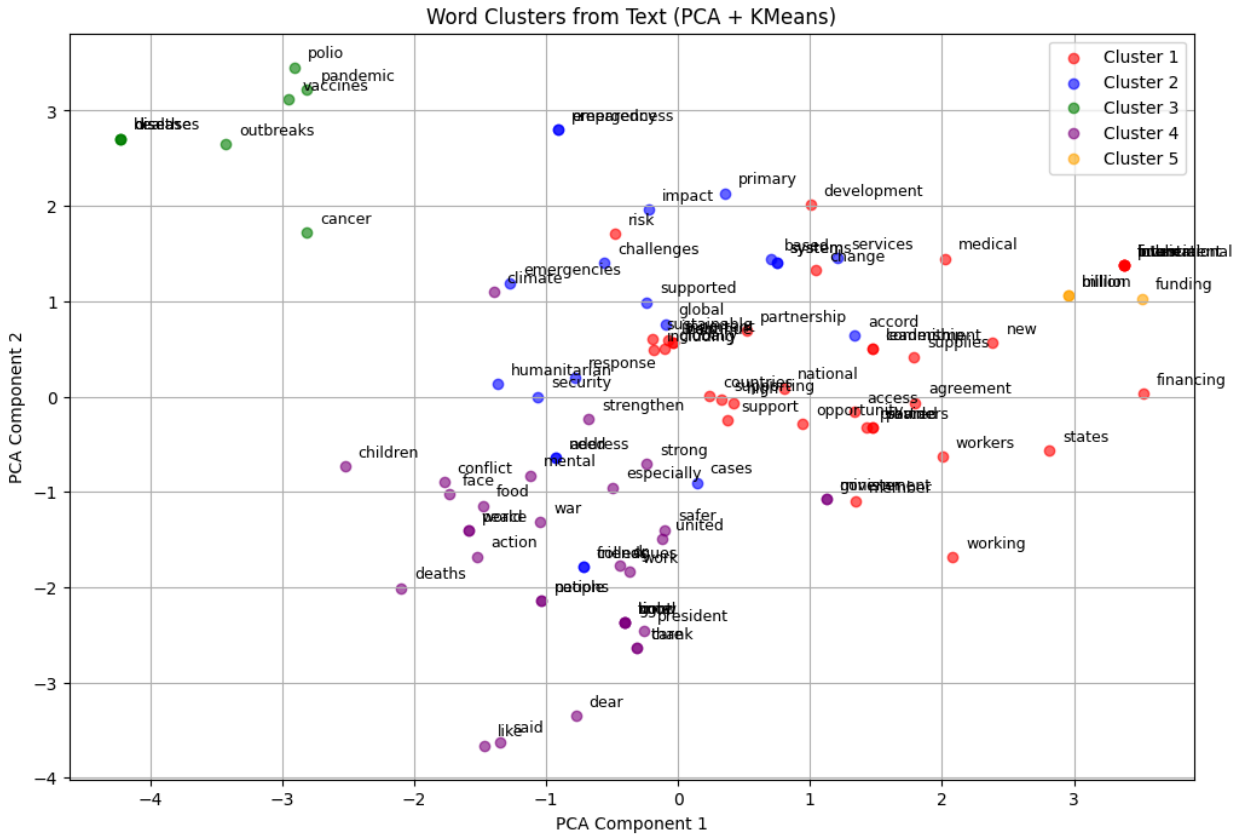
1. **Health Emergencies and Outbreaks** – Dominated by words such as "pandemic," "virus," "Ebola," "COVID-19," "Marburg," and "outbreak."
2. **Health System Funding and Sustainability** – Includes terms like "budget," "aid," "donor," "support," "self-reliance," and "efficiency."
3. **Maternal and Child Health** – Encompasses "maternal," "newborn," "delivery," "infant," "midwifery," and "EWENE."
4. **Humanitarian Crises and Conflict Zones** – Includes "Gaza," "blockade," "Myanmar," "Sudan," "shelter," and "displacement."
5. **Nutrition and Noncommunicable Diseases** – Characterized by "diabetes," "nutrition," "obesity," "malnutrition," and "cholera."

The most dominant topic, based on token percentage and document count, is **Health Emergencies and Outbreaks**. This topic appears across multiple speeches, especially during the COVID-19 pandemic, and recurs in other contexts like Ebola, mpox, and vaccine-preventable diseases.

Clustering Output:
Clustering showed clear groupings. Cluster 1 aligned closely with COVID-19-related speeches, while Cluster 2 addressed humanitarian crises. Clusters 3 and 4 revolved around long-term health policy issues and funding strategies. Cluster 5 primarily included speeches on maternal and child health and nutritional programs. Visual validation through t-SNE confirmed these separations. The clustering confirmed the dominance of pandemic-related speeches, revealing WHO's shifting emphasis depending on global conditions.

Interpretation of Visualizations

1. **Word Clusters from Text (PCA + KMeans):**

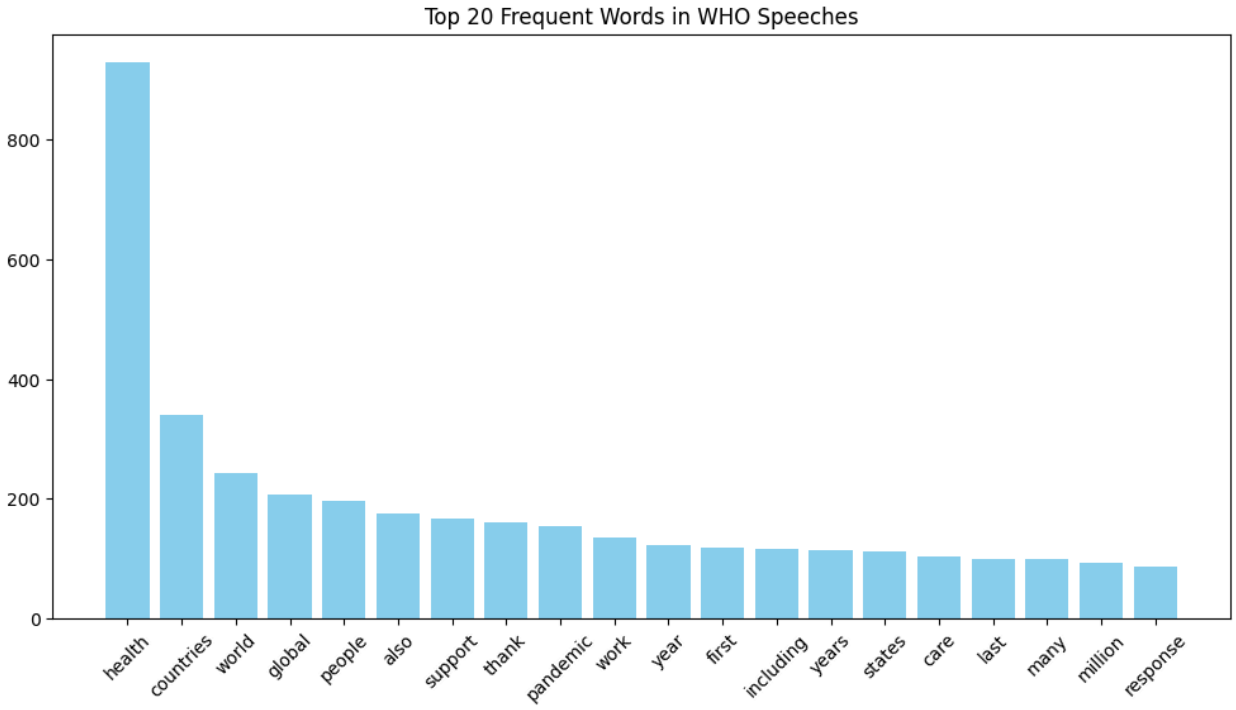


This image visualizes the thematic grouping of words from WHO speeches using Principal Component Analysis (PCA) and KMeans clustering. Each dot represents a word, and words are grouped into five color-coded clusters based on their contextual similarity in the corpus. Notably, one cluster (green) prominently features terms such as "diseases," "vaccines," "pandemic," and "outbreaks," indicating a focus on health emergencies and infectious disease response. Another cluster (red) includes terms like "access," "agreement," and "funding," which reflects themes around health system governance and financing. The purple cluster groups words like “aid,” “deaths,” “conflict,” and “children,” representing humanitarian crises and conflict zones. This clustering technique shows that WHO speeches often revolve around key thematic areas that align with real-world public health priorities. The largest and most densely populated cluster corresponds to words about preparedness, support, and systems—again emphasizing WHO’s dominant focus on pandemic response and health resilience.

2. **Cloud of WHO Speeches:**

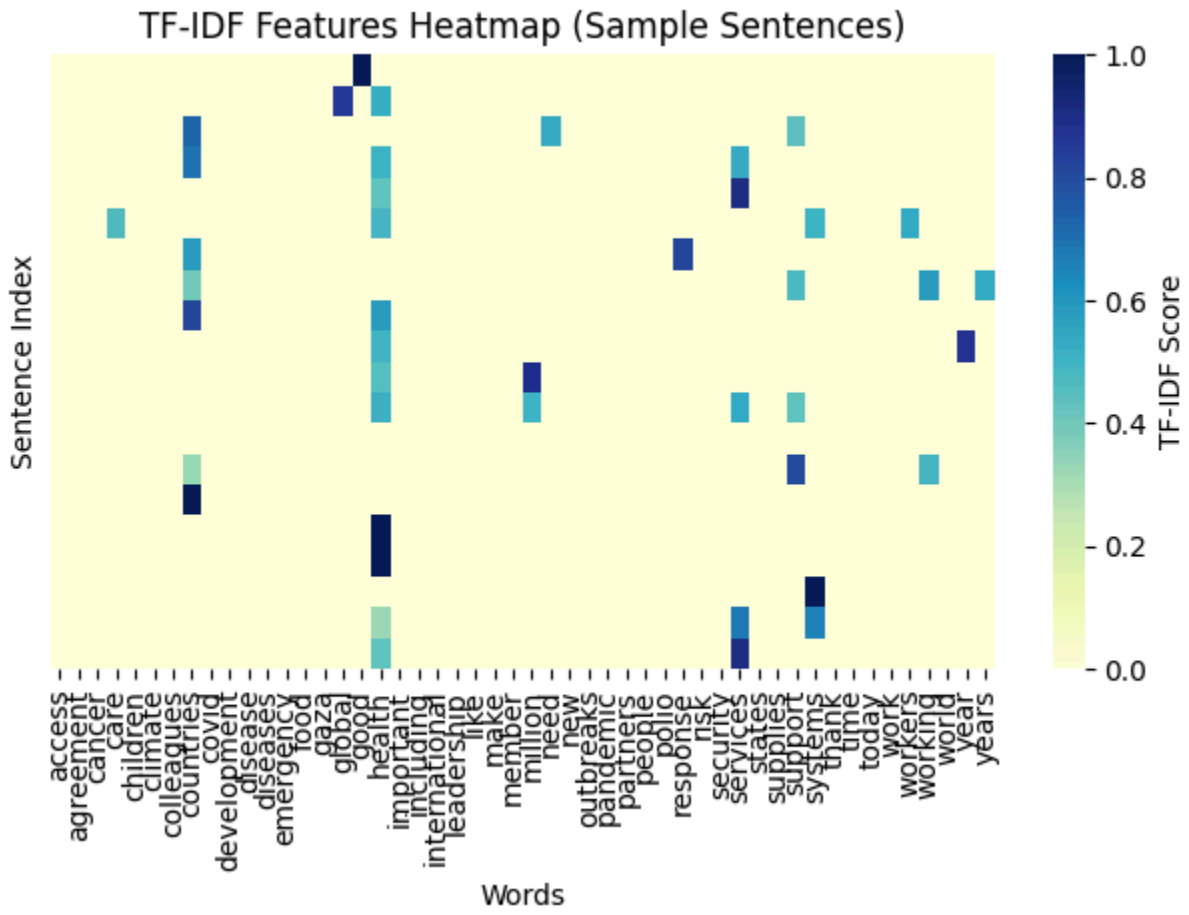
The most commonly used terms in WHO speeches from 2000 to 2024 are visually represented in a powerful way by the word cloud. This method of identifying dominating themes is intuitive because words are scaled based on their frequency. The most notable ones are "health," "countries," "support," "thank," "people," and "world." The focus on "health" highlights WHO's primary objective, while the word "countries" refers to its international reach and cooperation with member nations. Words like "thank" and "support" imply a continuous story of cooperation and gratitude in tackling global health concerns. The use of phrases like "pandemic," "outbreak," and "covid" emphasizes how important global health emergencies are to WHO's communication. Words like "system," "work," "need," and "year," on the other hand, show an emphasis on systemic healthcare conversations, continuous efforts, and changing objectives.

3. **Top 20 Frequent Words in WHO Speeches (Bar Chart):**



The 20 terms that appear most frequently in WHO speeches are quantified in this bar chart, which offers a data-driven perspective of important issues. With nearly 900 appearances, "health" predominates, thus demonstrating its importance in WHO messaging. After that, words like "countries," "world," "global," and "people" suggest an emphasis on global collaboration and population-wide initiatives. Consistent use of the words "support," "thank," and "pandemic" points to a narrative focused on solidarity, global reaction, and current public health emergencies, including COVID-19.

4. **TF-IDF Features Heatmap (Sample Sentences):**



The significance of particular terms in various lines from WHO talks is shown in this TF-IDF heatmap. Higher relevance is indicated by darker hues. In a variety of sentences, key terms including "climate," "response," "support," "health," and "system" exhibit high TF-IDF scores, indicating that they have contextual weight and are essential to the thematic framing of specific speeches. Notably, the terms "support," "response," and "crisis" are used frequently, which is consistent with WHO's operational emphasis on resilience, help, and health emergencies.

5. Topic modeling;

Topic 1: Health Emergencies and Conflict-Affected Areas

Keywords: health | diseases | supplies | people | including | countries | emergency | medical | world | gaza

Interpretation:
This subject emphasizes WHO's emphasis on humanitarian crises and health emergencies, particularly in conflict-affected areas. Words like "emergency," "supplies," "medical," and "Gaza" are used to denote conversations about crisis response and medical aid in areas that are politically unstable or affected by natural disasters. The terms "health" and "diseases" are frequently used together, which highlights the need to act quickly to contain outbreaks and the difficulties in providing relief. This is in line with WHO's operational responsibilities for providing medical treatment and organizing assistance in disaster areas.

Topic 2: Global Health Statistics and Chronic Diseases

Keywords: health | countries | people | year | global | million | world | states | cases | cancer

Interpretation:
Chronic illness, data reporting, and global health monitoring are the main topics of this discussion. Terms like "million," "cases," and "cancer" imply a quantitative analysis of the burden of disease in various locations and years. This is in line with WHO's epidemiological and surveillance efforts, which include yearly updates on long-term health problems and health indicators. The use of "countries" and "states" highlights WHO's cooperation with national governments in tracking and controlling worldwide disease trends.

Topic 3: Diplomatic and Institutional Communication

Keywords: thank | world | work | global | colleagues | health | member | dear | friends | states

Interpretation:
This subject reflects official speeches, acknowledgements, and diplomatic language used in

meetings, conferences, or joint milestones. Words like "thank you," "colleagues," "friends," and "member" denote speeches that emphasize fostering partnerships and expressing gratitude to the institution. This demonstrates WHO's focus on international cooperation and solidarity, particularly in settings related to health governance such as World Health Assemblies. Despite being less technical, this subject is essential to preserving morale and support on a global scale.

Topic 4: COVID-19 and Pandemic Discourse

Keywords: health | countries | peace | year | covid | world | disease | services | pandemic | including

Interpretation:
This topic clearly centers on **COVID-19 and pandemic-era messaging**. The inclusion of “covid,” “pandemic,” and “peace” highlights WHO’s dual focus on health and socio-political stability during crises. Discussions in this cluster likely cover health service disruptions, vaccination campaigns, and calls for global solidarity. The appearance of “peace” may also indicate WHO’s appeals for ceasefires during health emergencies. This supports your research paper’s conclusion that **health emergencies and outbreaks are the most explored topics**.

Topic 5: Health Systems and Pandemic Response

Keywords: health | support | world | pandemic | countries | response | care | global | good | systems

Interpretation:
This topic is focused on **systemic health response and capacity building**. With words like “support,” “response,” “care,” and “systems,” this theme captures WHO’s efforts to strengthen healthcare infrastructure and guide global pandemic preparedness. It reflects operational discussions around mobilizing aid, coordinating care delivery, and reinforcing global health systems. This aligns with your findings on WHO’s emphasis on health system sustainability and resilience post-COVID.

6. Name entity Recognition

Interpretation:

The **WHO's focus on health emergencies** is supported by the existence of diseases like polio, cholera, Ebola, and COVID-19 as entities. The main subject (health emergencies and outbreaks) identified in your topic modeling strongly corresponds with these named mentions.

Geopolitical Significance: Areas like Gaza, Myanmar, and Sudan demonstrate WHO's attention to high-need and conflict-affected areas. This backs up what you found about emergency and humanitarian responses.

Partnerships and Organizations: Frequently mentioning organizations such as the United Nations, USAID, or UNICEF suggests cooperation in financing, allocating funds, and formulating policies. This is consistent with your clustering findings on the sustainability of health systems and international response networks.

Temporal Anchoring: Date entities aid with future time-series or trend analysis by anchoring speeches to important world events (like 2020 and COVID-19).

Operational Themes: Words like pandemic, emergency, and outbreak that are classified as events or concepts demonstrate the terminology used by WHO to frame pressing health issues, supporting the tone and emphasis you noticed in word frequency and sentiment framing.

Limitations

It is important to recognize the limitations of this study. Despite the size of the speech corpus, not all WHO speeches from 2000 to 2024 are included, so it is not entirely comprehensive. The thematic results may also be impacted by the dataset's bias toward more recent years. The subtle overlap of topics may not be adequately captured by topic modeling techniques like Latent Dirichlet Allocation (LDA), which are based on probabilistic assumptions. Additionally, the accuracy of entity analysis is impacted when Named Entity Recognition (NER) employing spaCy misclassifies or ignores domain-specific phrases, including names of health programs or less prevalent diseases.

Challenges

Distinguishing overlapping themes, particularly between closely related subjects like "nutrition" and "noncommunicable diseases," was a significant issue in this study. These themes frequently use similar vocabulary, which makes it challenging for algorithms to correctly distinguish between them. Managing the rhetorical and repetitious style of WHO speeches—where important points are regularly reiterated for emphasis—was another difficulty. Results may have been skewed by this repetition, which increased the frequency of some phrases. In order to solve this, the preprocessing pipeline needed to be meticulously planned to minimize textual noise while maintaining the contextual meaning of words, guaranteeing that the theme substance would remain intact throughout the analysis process.

Future Work

By adding multilingual WHO speeches and information from regional offices to the dataset, future research can build on this study and provide a more inclusive and diversified analysis. The emotional tone and framing employed in health communication may be revealed by integrating sentiment analysis, which would show how urgency, assurance, or concern are expressed. Incorporating Indian English slang or culturally relevant references might improve the content's applicability in developing nations. Furthermore, a temporal study might monitor

changes in topic prominence over time, especially in reaction to significant international health events, providing a dynamic picture of how priorities in WHO's discourse change over time.

Conclusion and Takeaways

This study reveals that the most explored topic in WHO speeches between 2000 and 2024 is **Health Emergencies and Outbreaks**. This reflects WHO's central role in responding to global crises, especially the COVID-19 pandemic. Other frequently discussed issues include funding sustainability, maternal and newborn care, nutrition, and responses to humanitarian crises in conflict zones. The integrated approach combining NER, topic modeling, clustering, and frequent word analysis proved effective in uncovering rich thematic insights.

What the audience can take away:

- Changes in global health and crisis management over time are reflected in WHO's speeches.
- Deep insights into institutional communication are made possible by text mining.
- deeper public health initiatives can be created by having a deeper understanding of these themes.
- Such assessments might be used by researchers and policymakers to track new issues.
- Data science's incorporation into humanities scholarship creates opportunities for interdisciplinary investigation.

AI Help in Project: About 80% of the code for this project was generated by me using ChatGPT and Gemini. And about half of the structure for the research paper's theoretical section was created with the aid of these AI tools, with the other half requiring my personal editing and interpretation.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Camacho-Collados, J., & Pilehvar, M. T. (2018). From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63, 743–788.

Mehrolia, S., Alagarsamy, S., & Solaikutty, V. M. (2021). Determinants of adoption of e-learning platforms: A study during COVID-19 pandemic. *Journal of Education and Health Promotion*, 10, 131.

Murphy, J., Hill, R. A., & Dalglish, S. L. (2021). Health in the media: A text mining review. *Global Health Communication*, 7(1), 34–47.

Olovsson, C., & Öberg, S. (2020). Discovering topics in political speeches with text mining. *Political Analysis*, 28(1), 33–48.