Mining Insights from the United Nations General Assembly Debate Corpus

Prakhar Gupta
Rochester Institute of Technology
Rochester, United States of America
pg9349@rit.edu

Cecilia O. Alm
Rochester Institute of Technology
Rochester, United States of America
coagla@rit.edu

Abstract

Annually, the United Nations General Assembly meets to discuss global challenges and events occurring around the world. This meeting is characterized by speeches delivered by state representatives and delegates. This study explores both the outcomes from clustering of countries' speeches and thematic trends in speeches. Using word embedding representation, it analyzes past speeches and cluster them based on lexical semantic similarity. It also explores common themes through clusters of countries for certain years.

1 Introduction

At the annual General Assembly (GA) meeting of the United Nations (UN), speeches convey the nation's focus areas and agenda. Thus, UN GA speeches provide unique opportunities to explore geopolitical topics across the years. This study explores two questions:

1. What does the clustering of nations' speeches base on speech similarities reveal, and which countries' speeches share these similarities with one another?
2. Which topical trends in this corpus of speeches characterize particular nation-states, or can be generalized across nations?

The analysis focuses on large countries and the period from 1970 to 2019 to examine possible thematic patterns. Furthermore, it introduces an approach to group countries' speeches according to their similarities and topics discussed in the UN GA. To validate the insights gleaned, the analysis considers the coherence of the intra-cluster topics obtained. The results suggest that topics show reasonable affinity within clusters.

Using computational clustering can provide a tool to compare themes in speeches across several countries. This analysis also brings attention to the UN GA debate corpus to simulate future work.

1.1 Data

The corpus and metadata used in this study were obtained from the Harvard dataverse (Jankin Mikhaylov et al., 2017). This resource contains over 8000 speeches of various nation-states. Since the corpus covers 1970-2019, the temporal scope of the study was restricted to this time span.

1.2 Early Works

The United Nations General Assembly debate corpus (Baturo et al., 2017) offers speeches in machine-readable text format, combined with metadata per speech. The study presents an analysis of the speeches that leverage this corpus and analyzes multiple years of speeches while also analyzing specific countries' discussion themes from those speeches both in single and multiple dimensions.

Baturo and Dasandi (2017) combined speeches and identified over 15 topics across the corpus. Those 15 topics were analyzed and compared with each other to identify overlaps. Also, the study analyzed the relationship between World Bank's World Development Indicators(WDI) and the extent to which those topics are discussed by states. Another study Gurciullo and Mikhaylov (2017) used Doc2Vec to generate paired embeddings of countries and years and compared the representations for political notions such as *health, nuclear weapons, education*. The present work builds and extends on this work by using clustering to group countries and topic modeling over the clusters to recognize other themes.

In addition, Eckhard et al. (2021) considered topic modeling with speeches from the UN security council but is limited in its scope and source.

2 Structural and Thematic Trends

2.1 Word token analysis for 1970-2019

The corpus' texts were preprocessed with NLTK sentence and work tokenizers. Next, the tokens were grouped by year to analyze trends.

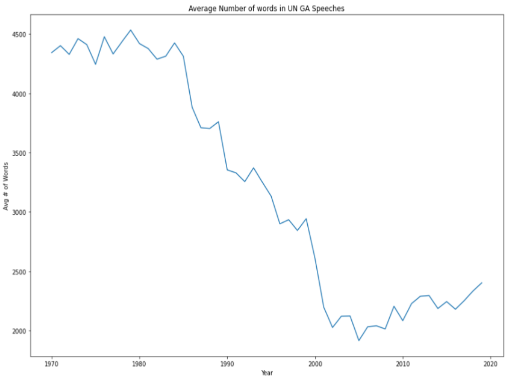

Figure 1: Analysing the average number of word tokens at UN GA speeches indicates that they have become shorter over time.

Fig 1. and Fig 2. both show a clearly decreasing trend for the number of word tokens from speeches in the UN GA between 1970-2019. For both words and sentences it appears they were reduced approximately by half.

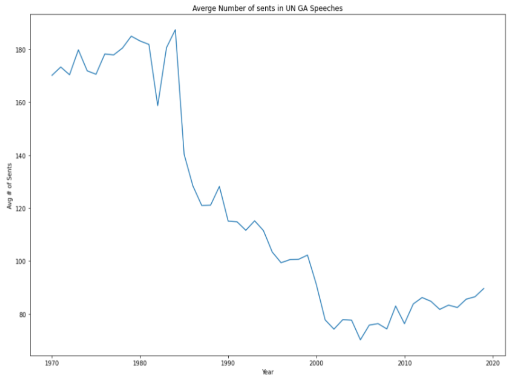

Figure 2: As expected, the number of sentences also decreases over time.

2.2 Sentiment in GA Speeches Associated with Major Countries

To explore how other nations characterize speeches, sentiment analysis was performed. In this analysis, a nation from the corpus was selected for analysis. Then all the speeches of this target country were removed from the corpus to ensure experimental soundness. GloVE embedding Pennington et al. (2014) helped identify word tokens associated with the selected country through similarity and these tokens formed capture groups. The remaining sentences from the corpus were analyzed with capture group tokens. The average sentiment associated with these sentences were grouped by year to analyze the sentiment across nations for a particular target.

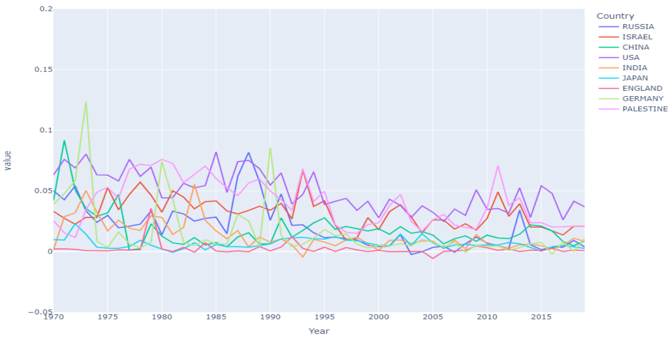

Figure 3: Sentiment associated with a subset of nine countries spanning years.

The analysis explored nine countries, shown in Fig 3. which shows how the sentiments of the world with respect to a target country have changed. Some countries like Japan and their capture group {japan, Japanese, tokyo} showed positive sentiment while countries like Germany {germany, german, berlin} extends toward both polarities. Most countries analyzed have an average positive sentiment throughout the corpus.

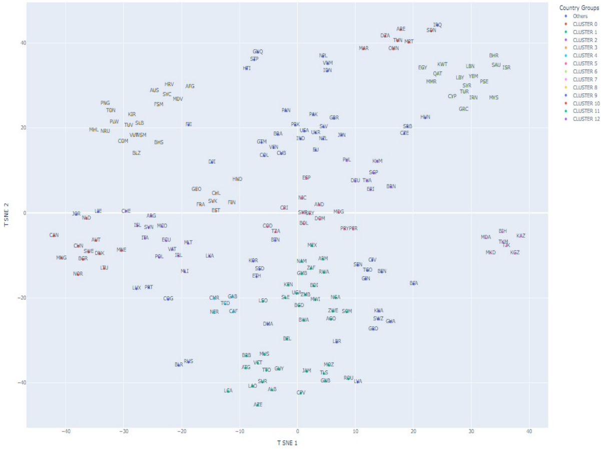

Figure 4: A cluster of countries based on applying speech embedding and dimensionality reduction with T-SNE, used Density Based Scan (Ester et al., 1996) uaes the approach to analyze 2019 UN GA speeches.

3 Identifying Topics Discussed for Individual Nation States

Next, the analyses extended to using the corpus and our approach to identify topics associated with or brought forth during the yearly UN GA debate.

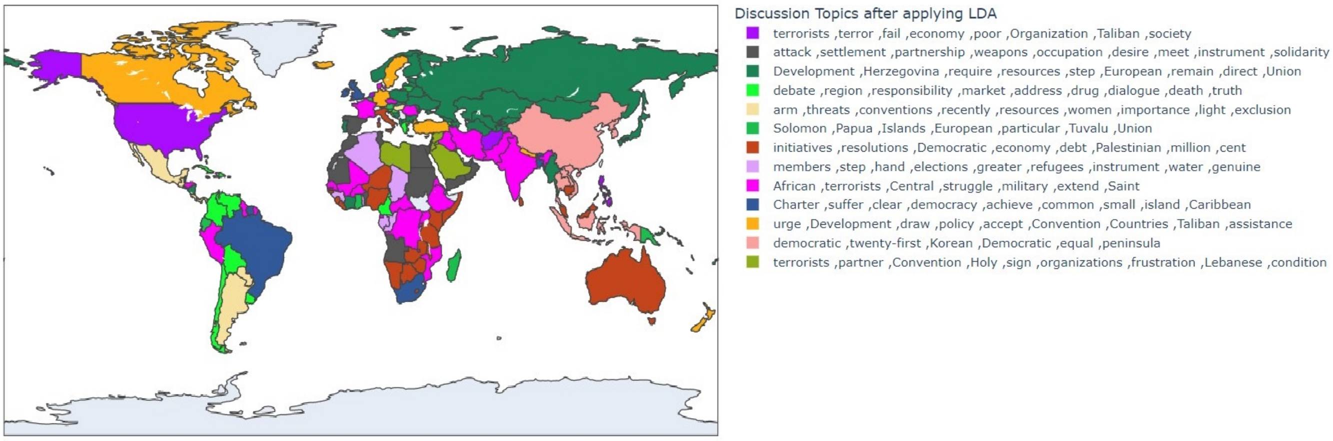
Discussion Topics after applying LDA
terrorists ,terror ,fail ,economy ,poor ,Organization ,Taliban ,society
attack ,settlement ,partnership ,weapons ,occupation ,desire ,meet ,instrument ,solidarity
Development ,Herzegovina ,require ,resources ,step ,European ,remain ,direct ,Union
debate ,region ,responsibility ,market ,address ,drug ,dialogue ,death ,truth
arm ,threats ,conventions ,recently ,resources ,women ,importance ,light ,exclusion
Solomon ,Papua ,Islands ,European ,particular ,Tuvalu ,Union
initiatives ,resolutions ,Democratic ,economy ,debt ,Palestinian ,million ,cent
members ,step ,hand ,elections ,greater ,refugees ,instrument ,water ,genuine
African ,terrorists ,Central ,struggle ,military ,extend ,Saint
Charter ,suffer ,clear ,democracy ,achieve ,common ,small ,island ,Caribbean
urge ,Development ,draw ,policy ,accept ,Convention ,Countries ,Taliban ,assistance
democratic ,twenty-first ,Korean ,Democratic ,equal ,peninsula
terrorists ,partner ,Convention ,Holy ,sign ,organizations ,frustration ,Lebanese ,condition

Figure 6: Major discussion topics in the speeches of each country from the 2001 UN GA session organized by their respective cluster.

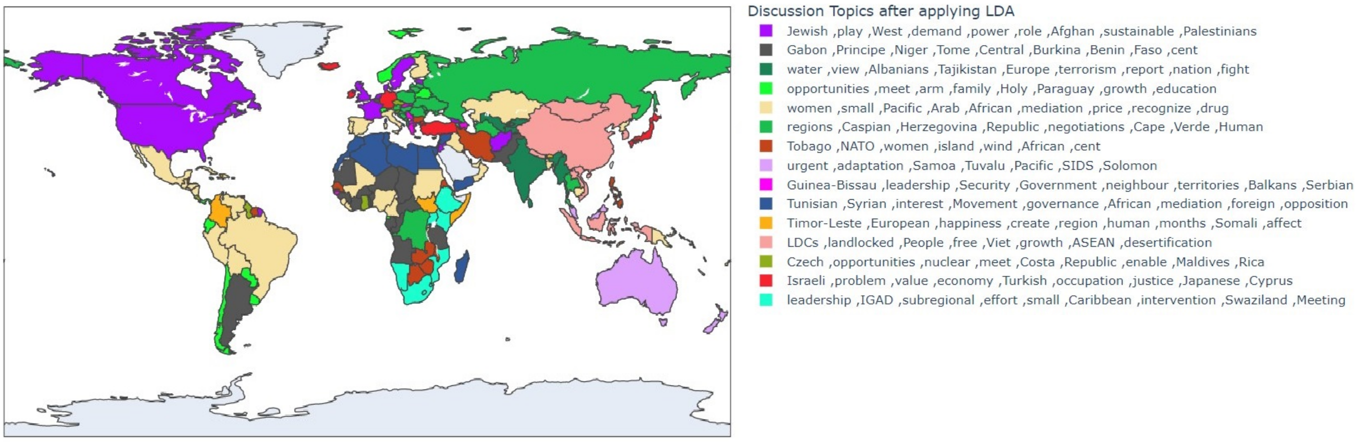
Discussion Topics after applying LDA
Jewish ,play ,West ,demand ,power ,role ,Afghan ,sustainable ,Palestinians
Gabon ,Principe ,Niger ,Tome ,Central ,Burkina ,Benin ,Faso ,cent
water ,view ,Albanians ,Tajikistan ,Europe ,terrorism ,report ,nation ,fight
opportunities ,meet ,arm ,family ,Holy ,Paraguay ,growth ,education
women ,small ,Pacific ,Arab ,African ,mediation ,price ,recognize ,drug
regions ,Caspian ,Herzegovina ,Republic ,negotiations ,Cape ,Verde ,Human
Tobago ,NATO ,women ,island ,wind ,African ,cent
urgent ,adaptation ,Samoa ,Tuvalu ,Pacific ,SIDS ,Solomon
Guinea-Bissau ,leadership ,Security ,Government ,neighbour ,territories ,Balkans ,Serbian
Tunisian ,Syrian ,interest ,Movement ,governance ,African ,mediation ,foreign ,opposition
Timor-Leste ,European ,happiness ,create ,region ,human ,months ,Somali ,affect
LDCs ,landlocked ,People ,free ,Viet ,growth ,ASEAN ,desertification
Czech ,opportunities ,nuclear ,meet ,Costa ,Republic ,enable ,Maldives ,Rica
Israeli ,problem ,value ,economy ,Turkish ,occupation ,justice ,Japanese ,Cyprus
leadership ,IGAD ,subregional ,effort ,small ,Caribbean ,intervention ,Swaziland ,Meeting

Figure 7: Major discussion topics in the speeches of each country from the 2011 UN GA session organized by their respective cluster.

3.1 Clustering Based on Speech Similarity

Fig 5. show the result after vector embedding from speeches are combined with a density-based scan algorithm (Ester et al., 1996) to cluster countries with similar speeches. For this step, Spacy NER (Honnibal and Montani, 2017) was used to remove *geopolitical entities*, [1] from the speeches to avoid them being assessed as similar due only to GPE cross-references. Vector embeddings were created for the speeches using gensim's Doc2Vec (Le and Mikolov, 2014). These embeddings were compressed into two dimensions with scikit-learn's T-SNE(van der Maaten and Hinton, 2008) and substantial variations were passed to scikit-learn's DBSCAN (Ester et al., 1996) to obtain clusters.[2]. Clustering was evaluated using a silhouette score. With respect for Fig 5. the score was 0.4003 which indicates that the model does not overlap in clusters or mislabeled data points.

3.2 Notable Themes in Clusters

To explore discussion themes multiple iterations were analyzed. One iteration included speeches in a cluster combined to form a list of documents where each involved a country's speech for a given year the clusters were obtained before. In pre-processing, both most frequent stop words and frequent generic words such as *secretary-general, president, convey* etc. were removed from the corpus geo-Political entities were not removed to allow insights about any meaningful topics in this step.Gensim's LDA (Hoffman et al., 2010) was used to identify three prominent topics for each

[1] Cluster after removal of stop words offered similar results.

[2] K-means and hierarchical clustering using scikit-learn were also explored but these approaches are sensitive to outliers and were dropped.

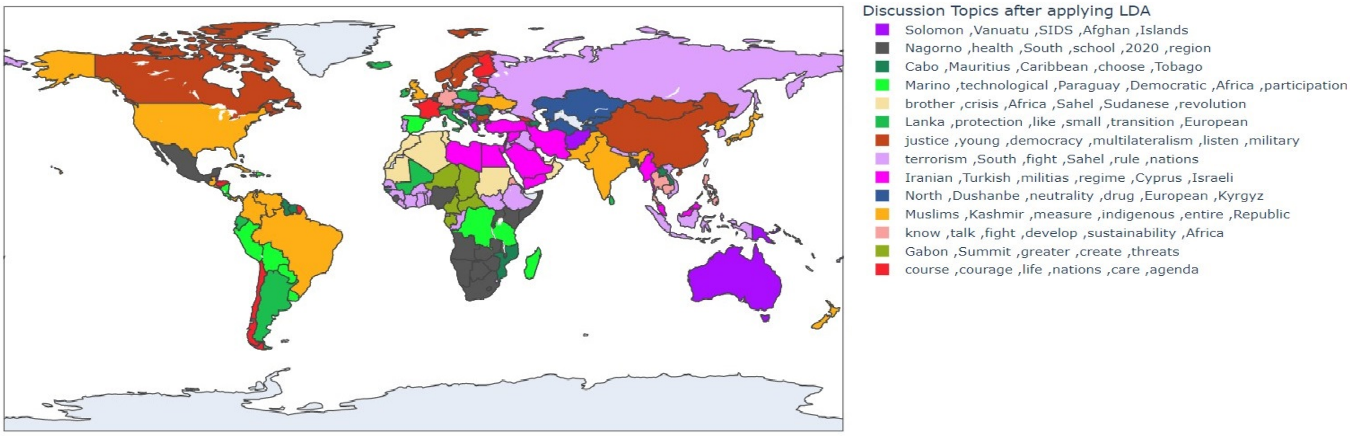
Discussion Topics after applying LDA
Solomon ,Vanuatu ,SIDS ,Afghan ,Islands
Nagorno ,health ,South ,school ,2020 ,region
Cabo ,Mauritius ,Caribbean ,choose ,Tobago
Marino ,technological ,Paraguay ,Democratic ,Africa ,participation
brother ,crisis ,Africa ,Sahel ,Sudanese ,revolution
Lanka ,protection ,like ,small ,transition ,European
justice ,young ,democracy ,multilateralism ,listen ,military
terrorism ,South ,fight ,Sahel ,rule ,nations
Iranian ,Turkish ,militias ,regime ,Cyprus ,Israeli
North ,Dushanbe ,neutrality ,drug ,European ,Kyrgyz
Muslims ,Kashmir ,measure ,indigenous ,entire ,Republic
know ,talk ,fight ,develop ,sustainability ,Africa
Gabon ,Summit ,greater ,create ,threats
course ,courage ,life ,nations ,care ,agenda

Figure 8: Example discussion topics with their most frequent word tokens in the speeches of each country from the 2019 UN GA session organized by their respective cluster.

cluster. [3] These iterations were repeated for all clusters obtained hence identifying topics spanning countries in a given UN GA session.

4 Additional Analysis Results

Fig 7. shows that countries discussed different topics pertaining to their situation, for example *SIDS* and *Solomon Islands* for Solomon (SOL) and New Zealand (NZL).

Fig 8. shows a cluster of countries for the 2019 UN GA. Speeches 'topics appeared to be linked the following year. For example, topics like *Iranian, Turkish, militias, regime, Cyprus, Israeli* appeared in a Middle Eastern cluster based on the Geo-political domain knowledge one would expect. It's interesting to see Middle Similar results could also be observed for Oceania, Southern Africa, and Central Asia clusters.

Fig 9. indicates substantial coherence (Röder et al., 2015) for the clusters. The lowest coherence is for the *Others* country cluster which includes countries whose speeches were generally dissimilar.

5 Discussion and Conclusion

The clustering and topic model analyses on UN GA speeches suggested that countries sharing geography or geo-political concerns tended to cluster together. This study also attempted to bring attention to this underused corpus in research efforts.

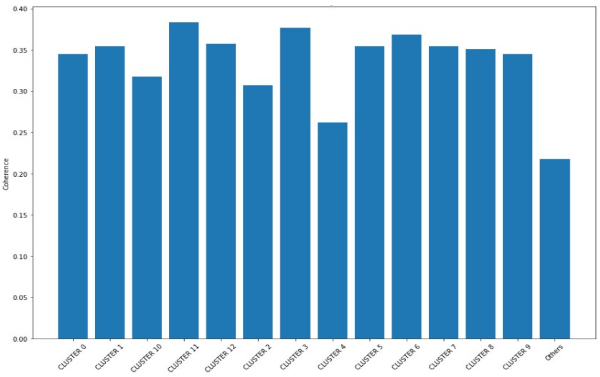

Figure 9: Coherence metric for topics obtained for clusters using DBSCAN (Ester et al., 1996)

References

New zealand united nations general assembly fifty-sixth session general debate statement by the honourable Phil Goff, minister of foreign affairs and trade of new zealand, monday 12 november 2001.

Solomon islands statement by the honourable Manasseh Sogavare, prime minister of solomon islands to the general debate of the fifty-sixth session united nations general assembly, 10 november 2001.

Alexander Baturo and Niheer Dasandi. 2017. What drives the international development agenda? An NLP analysis of the united nations general debate 1970–2016. In *2017 International Conference on the Frontiers and Advances in Data Science (FADS)*, pages 171–176.

Alexander Baturo, Niheer Dasandi, and Slava J. Mikhaylov. 2017. Understanding state preferences with text as data: Introducing the UN General Debate corpus. *Research & Politics*, 4(2):2053168017712821. _eprint: https://doi.org/10.1177/2053168017712821.

Steffen Eckhard, Ronny Patz, Mirco Schönfeld, and

[3] Similar results were obtained for 1991 and 1981 in comparing analysis.

Hilde van Meegdenburg. 2021. International bureaucrats in the UN Security Council debates: A speaker-topic network analysis. *Journal of European Public Policy*, 30:1–20.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Knowledge Discovery and Data Mining*.

Stefano Gurciullo and Slava Mikhaylov. 2017. Detecting Policy Preferences and Dynamics in the UN General Debate with Neural Word Embeddings. ArXiv:1707.03490 [cs, stat].

Matthew Hoffman, Francis Bach, and David Blei. 2010. Online Learning for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Slava Jankin Mikhaylov, Alexander Baturo, and Niheer Dasandi. 2017. United Nations General Debate Corpus. Technical report, Harvard Dataverse.

Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1188–1196. PMLR. ISSN: 1938-7228.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 399–408, New York, NY, USA. Association for Computing Machinery.

Laurens van der Maaten and Geoffrey Hinton. 2008. Viualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.