Mining Insights from United Nations General Assembly Debate Corpus

Prakhar Gupta
ENGL.681 Natural Language Processing I
Rochester Institute of Technology
Rochester, United States of America
pg9349@rit.edu

***Abstract*—Every year in the month of September - October one of the biggest global bodies, the United Nations General Assembly, meets to discuss and debate on the global challenges and events that happened or are happening around the world.This event consists of speeches by various states representatives and delegates. Hence it becomes imperative to develop an approach to identify the trends and create an approach which is both granular in identifying underlying topics of each nation state is generalized to identify countries which also share the same topics in their respective speeches. This paper does the above by analyzing past speeches to mining relevant insights and also proposes a novel approach to cluster countries together based on speech similarity and identify the common topics discussed by each cluster of countries for the respective year.**

***Index Terms*—United Nations, topic modelling, document similarity, GPE sentiment analysis**

I. Introduction

The United Nations consists of 193 sovereign states and is the world's largest international body.This provides a unique set of opportunities and challenges whereas on one side such event provide one of the most pristine and rich textual data and on the other side there are plethora of topics discussed during each years meet amounting to speeches which address array of interests and agenda.

This paper analyzes the sentiments of United Nations member states for major countries over a time period of 1970-2019 identifying the pattern. Furthermore this paper introduces a novel approach to group countries according to their speech similarities and within each group identify the topics which are discussed and put forth by states of the general assembly. To validate the results of the approach I have used evaluation metrics, coherence on the intra cluster topics obtained and the topics do show a reasonable affinity with the associated clusters and associated countries.However to limit the scope I haven't tried to reason the trends or insights which are observed and hypothesis causation for the insights.

A. Data

The data set and meta data used for this study is obtained from Harvard dataverse [2] which contains approximately 8000+ speeches of various nation states from the year of 1970-2020. Since the metadata covers 1970-2019 I have restricted the scope of the study from 1970-2019.

B. Early Works

The United Nations General Assembly debate corpus was first introduced by [1] "Understanding state preferences with text as data: Introducing the un general debate corpus"(A. Baturo, N. Dasandi, and S. J. Mikhaylov) paper does a great job by curating and serializing the speeches from different formats to machine readable text format and providing associated metadata with each speech. The paper [1] goes ahead with analyzing the speeches first in a single dimension using "Wordscore" [3](Laver, Benoit and Garry, 2003) to draw an empirical score for the topic of USA-Russia rivalry in world politics. The paper then goes ahead analyzing the speeches in multi dimensions using correspondence analysis. Even though the paper does bring forth valuable insight it fails to develop a holistic approach in analyzing yearly speeches and only address a handful of countries' political standings across time.

The paper [4] "What drives the international development agenda? An NLP analysis of the United Nations general debate 1970–2016"(A. Baturo and N. Dasandi) builds upon the existing research to agglomerate speeches to identify 16 topics across the breadth of corpus. This approach though exposes the underlying topics discussed at United Nations, fails to account for the fact of the sheer number of countries in the corpus making the approach too general for identifying any country or year specific topics.

II. Changing Trends

A. Tokens across 1970-2019

As part of Data processing and Exploratory data Analysis on the corpus. Sentence tokenizer and work tokenizer from NLTK was used to split each speech into sentences and words. This was done as pre processing and also to estimate year on year trends

The Fig 2. and Fig 3. both shows a overall decreasing trend in the number of words spoken in the UN and is both cases we observe that between 1970-2019 both # of words and sentences has reduced by half.

B. Sentiments for Major Countries

"How does the world view a country?" To answer this question sentiment analysis was performed. This was achieved in multiple iterations, in every iteration a country was selected as a target country and speeches of that target country were

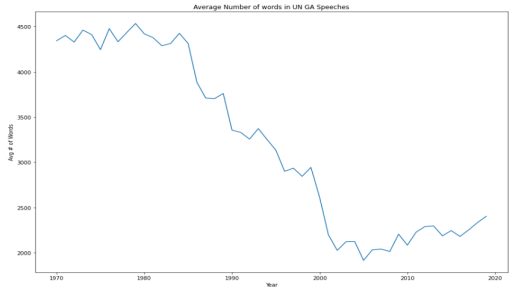

Fig. 1. Average Number of Words at UN GA.

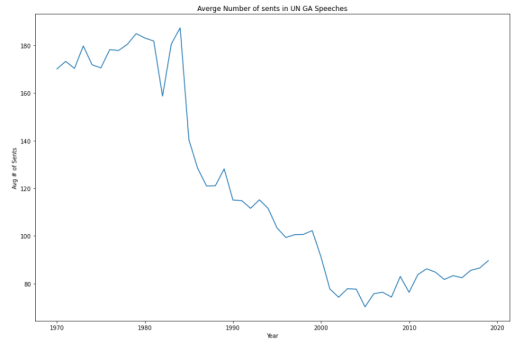

Fig. 2. Average Number of Sentences at UN GA.

removed from the corpus. Glove Embedding [5] was used to identify the associated words to the target country and these words formed a set of capture group. The whole corpus minus the target country was passed to identify the sentences across which the capture group words appear. These sentences were then passed through Spacytextblob sentiment analyzer to identify sentiments of each country against target.

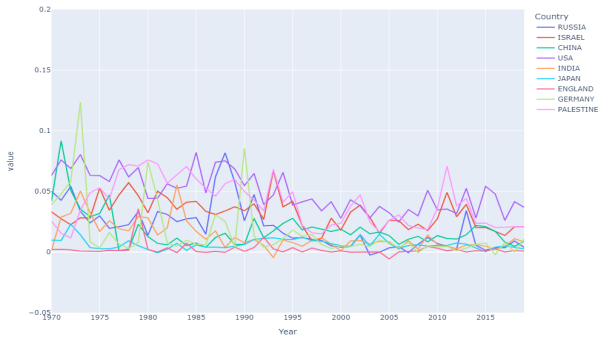

Fig. 3. Sentiment of the world with respective to Major Countries

To limit the scope of this study, iterations were run across 9 major countries. Fig 4. shows how the sentiments of the world with respect to a target country have changed. A trend could be observed where some countries like Japan and by associated Japan's capture group {'japan','japanese','tokyo'} have a almost constant slight positive sentiment while countries like Germany and its capture group {'germany','german','berlin'} have its high and lows. One interesting observation is that mostly all the countries analyzed for this study have an average positive sentiment throughout the corpus.

III. Identifying Topics Of Interest For Each Nation State

To identify the topics discussed by each nation during each year's General Assembly meet.We propose a novel approach, unlike previous works on this corpus this approach is granular to identify the topics which each nation brings forth and also help in identifying the countries which bring forth the same topics.It consist of multiple iterations in which in each iteration every year speeches are passed to steps given below [1]

A. Clustering of Nations based on Speech Similarity

Fig 5. show result after generating vector embedding for speeches and running them across a Density based scan algorithm to cluster countries which have similar speeches

Fig. 5. Cluster of Countries based on speech embedding, Dimensionality reduction using T SNE and clustering using Density Based Scan for 2019 UN GA Speeches

As a preprocessing for this step, Spacy NER [6] was used to remove all GPE(GeoPolitical Entities) [2] from the speeches to avoid speeches being similar due to cross reference. Vector embeddings were created for these speeches using Gensim Doc2Vec [7]. These embedding were compressed into two dimensions using scikit-learn T-SNE [8] and the dimensions of maximum variations were passed to DBSCAN [9] to generate clusters [3].

B. Major Topics In Each Cluster

To identify discussion topics multiple iterations were run. Each iteration represents speeches in a cluster obtained from the earlier step.Stop Words and words common at UN GA were removed. Gensim LDA [10] was used to identify 3 prominent topics in each iteration hence for each cluster.

[1]Due to limitations plots shown are for 2019 but similar results were obtained for other years showcasing the generality of the approach

[2]Cluster after removal of stopwords also gave similar results

[3]KMeans and Hierarchical Clustering were also tried but since these approaches are sensitive to outlier they were dropped

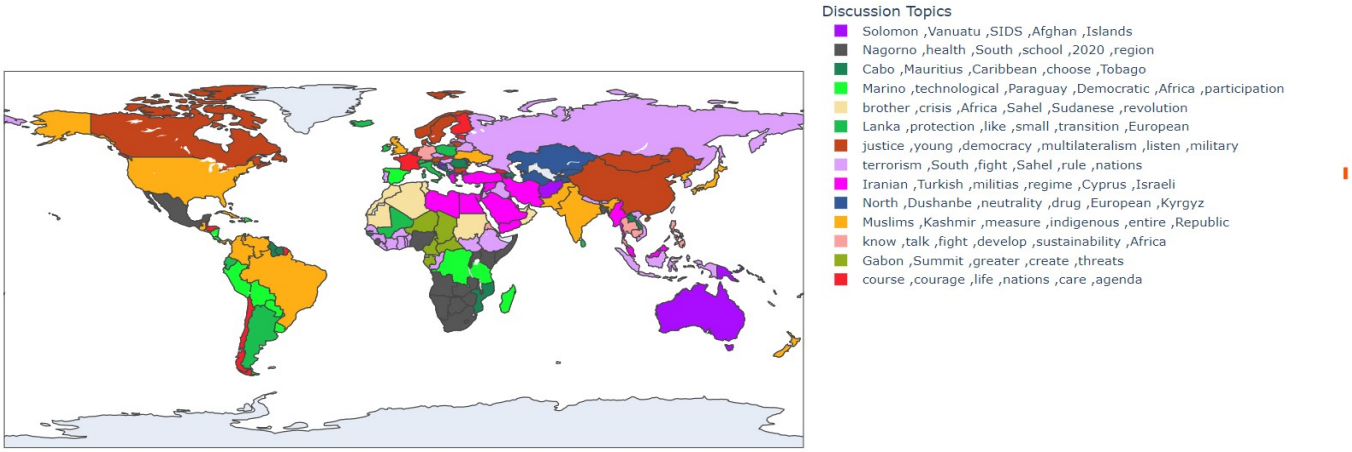

Fig. 6. Major discussion topics in the speeches of each countries at 2019 UN GA session organised by their respective cluster

IV. Results

Fig 6. shows major topics for each cluster of countries for 2019 UN GA.Couple of insights can be drawn from the figure.The topic appears to be very specific to the country or the geography or based on events of the following year. For example topics like 'Iranian ,Turkish ,militias ,regime ,Cyprus ,Israeli ' are very specific to the Middle east country cluster which is based on Geo-political domain knowledge one would expect. It's really interesting to see that contrary to popular belief, middle east countries' speeches are almost identical to one another. Similar results could be observed in Oceania, Southern African and central Asia clusters.

Fig 7. act as a validation for the approach.We do achieve significant coherence for individual cluster.Also the lowest coherence is for the "Others" country cluster which includes countries whose speeches were not similar to any group. This again validates our approach.

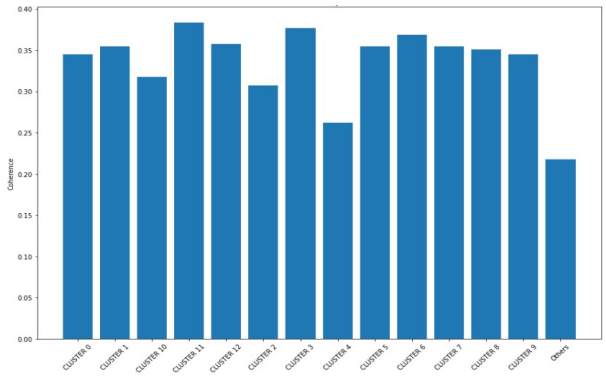

Fig. 7. Coherence metric for Cluster of Countries

V. Discussion and Conclusion

Since the data consist of speech of political entities the use of this data is completely ethical. All sources are cited as needed. We have tried our best to evaluate the approach using NLP evaluation metrics but true evaluation can be achieved only through domain experience in Geo-politics. To limit the scope we only analyzed speeches in the later years of corpus. This paper is also an attempt to bring this corpus to limelight so more research could be done on the same.

Acknowledgment

I would like to thank Professor Alm for her continuous guidance and my peers for helping me brainstorm ideas for this paper

References

[1] A. Baturo, N. Dasandi, and S. J. Mikhaylov, "Understanding state preferences with text as data: Introducing the UN General Debate corpus," Research Politics, vol. 4, no. 2, p. 2053168017712821, 2017, doi: 10.1177/2053168017712821.

[2] S. Jankin Mikhaylov, A. Baturo, and N. Dasandi, "United Nations General Debate Corpus," Harvard Dataverse, 2017.

[3] M. Laver, K. Benoit, and J. Garry, "Extracting Policy Positions from Political Texts Using Words as Data," The American Political Science Review, vol. 97, no. 2, pp. 311–331, 2003.

[4] A. Baturo and N. Dasandi, "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), Oct. 2017, pp. 171–176. doi: 10.1109/FADS.2017.8253221.

[5] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014, pp. 1532–1543. doi: 10.3115/v1/D14-1162.

[6] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017.

[7] Q. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," in Proceedings of the 31st International Conference on Machine Learning, Jun. 2014, pp. 1188–1196. Accessed: Dec. 08, 2022. [Online]. Available: https://proceedings.mlr.press/v32/le14.html.

[8] Van der Maaten, L.J.P.; Hinton, G.E. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9:2579-2605, 2008.

[9] Ester, M., H. P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press, pp. 226-231. 1996.

[10] R. Rehurek and P. Sojka, "Gensim–python framework for vector space modelling," NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, vol. 3, no. 2, 2011