Analyzing and Contextualizing UN General Assembly Speeches
(MDA Team Armenia)

David, Daniel, Khachatur, Christine, Nadim, Luca

May 31, 2021

Introduction

This project pertains to a collection of transcripts of UN speeches from 1970 to 2018. The project aims to apply several Natural Language Processing and Machine Learning techniques to identify the most frequently discussed UN topics over the years. Similarly, the project attempts to show how countries' stances change have evolved in regards to these subjects and corroborate the yielded results with external data. The final result is an application that showcases all the work done in a user-friendly and interactive manner.[1]

Substantively speaking, we conduct topic modelling and sentiment analysis on the abovementioned dataset to derive key trends in UN general assembly speeches. We then look at two of the most hotly-debated topics in the assemblies—namely sustainable development and the Arab-Israeli conflict—as case studies to further analyze and contextualize our initial, general findings.

Preparing The Data

Prior to conducting any potent analysis, though, the data had to be cleaned. For this, text files containing speeches for all years and speakers had to be merged. The result is one dataframe with each speech as a row. The year on which the speech was pronounced, as well the country, are variables of interest, in addition to the raw text itself. Additionally, this last piece of the data had to be further cleaned for topic modelling to be performed on the dataset. To achieve this, we create a class labelled DataCleaners to clean the text using regular expressions, remove various stop words and lemmatization. It should be noted that in addition to the various stop words removed by NLTK, we created a list of stopwords which we construct by generating a list of some of the most commonly reoccurring and undesirable bigrams over the years. Thus, phrases such as "united nations" are deleted from the dataset, as they provide no particular substantive insight.

Topic Modelling Via LDA

We begin modelling topics by visualizing the frequencies of the most common bigrams in our cleaned dataset. A snapshot of the histogram shown in our web application.

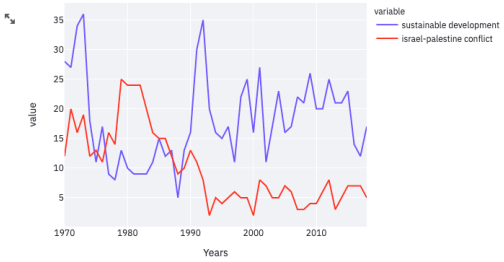
We supplement this preliminary analysis by identifying key topics in our dataset with the use of a Latent Dirichlet Allocation (LDA) machine learning model. This way, we cluster salient terms in the data, and can interpretively assign a topic to each filing. An intertopic map of the eight topics is depicted in the app, with the frequencies of the most relevant words for each topic displayed interactively on the right hand side. We assign the topics as shown in the table below.

Having determined these topics, we visualize the evolution of two topics in particular, which garner our attention later as mentioned in the introduction. Sustainable development and the Arab-Israeli Conflict have, as topics, followed different trends in terms of attention at the general assembly (see graph below). While the number of speeches centered on sustainable development has been quite

[1] Link to the App: https://share.streamlit.io/dbejarano31/team_armenia_mda/main/app/MDA-appy.py.

Topic	Topic Name
1	Development of Africa
2	Human Rights
3	International Security
4	Nuclear Politics
5	Economic Development
6	Arab-Israeli Conflict
7	World Peace
8	Sustainable Development

volatile until 2018, the topic has remained fairly popular among assembly speakers. On the other hand, attention for the Arab-Israeli conflict has significantly dropped over the years. This year's flare up in tensions in Gaza may, however, bring the topic back to the international arena.



Sentiment Analysis

After the main topics were extracted in the previous step, further investigation into the stances of countries on different topics was required. To glean the most out of the data, sentiment analyses were conducted using two different lexicons, Textblob and Vader. Textblob retrieved the polarity score and the subjectivity score, while Vader returned the positivity score, the neutrality score and the negative score. A brief definition of each of the said scores is given in the table below. These scores thus provide insight for each individual speech over the years, and given that the main topic of each speech was identified in the previous step, it became much easier to diagnose the feelings of countries toward a certain topic and how they evolve over the years. The app that was created in the scope of this project can help an end user further investigate these aspects with a few clicks.

Score	Definition
Polarity	Shows the general sentiment in the speech. It is computed from the positive, negative and neutral scores.
Subjectivity	Measures the degree to which subjective emotions are expressed in the speech.
Positivity	A score determined based on the degree to which positive words are used.
Negativity	Similar to positivity, but for negative words.

Case Studies

Having performed these analyses, we try to contextualize and explain some of the observed phenomena by referring to further datasets on two of the identified topics: sustainable development and the Arab-Israeli conflict. We choose to zero-in on the former due to its salience in our analyses and because climate change is the principal scope of this project. We also further look at the latter due to its topical nature, in light of recent developments in the region.

Case 1: Analyzing and Mapping Sustainability Discourse

Here, we try to assess how countries' discourse on sustainable development is reflected in their own sustainability policies. For this, we refer to the Sustainable Development Index (SDI) developed by economic anthropologist Jason Hickel [2] [3] Using this data, we construct ratio which we call Honesty

[2] About: https://www.sustainabledevelopmentindex.org/about.
[3] Methods: https://www.sustainabledevelopmentindex.org/methods.

Ratio (HR). The honesty ratio is constucted as follows:

$$HR = \frac{Normalized\ topic\ count}{Mean\ SDI},$$

where the normalized topic count is the number of times a particular country has alluded to sustainable development between 1990-2018 normalized to be comprised in the range [0,1], and the denominator is the average SDI value for the country over the same time period. Thus smaller HR values indicate stronger speech-to-implementation performance. We map these values in our app. Some of the worst performers are displayed in the table below.

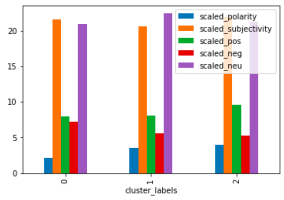
Country	Count	Mean SDI	Honesty Ratio
USA	26	0.231	4.324
Canada	10	0.289	1.246
United Kingdom	15	0.481	1.163

Case 2: Analysis of Arab-Israeli Conflict

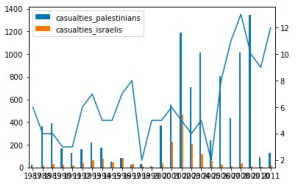
In order to retrieve some additional insights from the Arab-Israeli topic, we decided to use the sentiment score to cluster the countries. Assessing the amount of clusters is often tricky. An obvious choice is to cluster the countries into 2 clusters, however there were indications that 3 clusters, and even sometimes 4 were possible.

For this assignment we opted for the K-Means clustering algorithm. Essentially, we will use all five scores yielded in the Sentiment Analysis phase. As a further step, the clustering was repeated over 10-year periods to see if there were countries that changed their stances over the years. For the k-means clustering, k was chosen to be 3 to facilitate interpretation. The underlying assumption is that we encounter countries that are Palestine supporters, Israel supporters and which ones are neutral on this topic.

As you can observe below, cluster 0 has the lowest polarity, very high subjectivity and the highest negativity, which means countries in this cluster are likely Israel supporters, for example USA is in cluster 0. Cluster 1 seems to be quite neutral, possessing the lowest subjectivity score and highest neutrality score. Lastly, cluster 2 has the highest polarity and positivity and is quite subjective, most Arab countries were found to be in cluster 2. All these findings further prove the quality of the conducted Sentiment Analysis.



The number of UN speeches pertaining to the Arab-Israeli conflict has significantly decreased since 1980. It should be noted that this could change for years to come, with an uptick in tensions in the region. However, as can be shown in the figure below (based on data from BtSelem[4]), UN speeches are fairly uncorrelated to casualties and damage in the Palestinian territories.



[4] https://www.btselem.org/statistics